Entdeckung von Täuschung:

Von Alltagsvorstellungen zu empirisch fundiertem Wissen

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Philisophie des Fachbereiches 06

der Justus-Liebig-Universität Gießen

Maike Miriam Breuer

aus

Neuss

2008

Dekan:	Prof. Dr. Joachim C. Brunstein
1. Berichterstatter:	Prof. Siegfried L. Sporer, Ph.D.
2. Berichterstatter:	Prof. Dr. Joachim Stiensmeier-Pelster

DANKSAGUNG

Abschließend möchte ich all jenen danken, die mich bei der Fertigstellung dieser Arbeit unterstützt haben, insbesondere meinem Betreuer Prof. Siegfried L. Sporer, Ph.D., für seine hilfreichen fachlichen Rückmeldungen. Bei meinem Zweitgutachter Prof. Dr. Joachim Stiensmeier-Pelster möchte ich mich ebenfalls für sein Interesse an dieser Arbeit recht herzlich bedanken.

Zudem danke ich Dr. Marc-André Reinhard für seine Anregungen zur ersten Studie. Auch Anja Balser, Agnes Burek, Sonja Fischer, Johanna Hermann, Carolin Kaufmann, Lisa Kocher, Anne Kreh, Florian Lautner, Kathrin Lutz, Patricia Mathyl, Bettina Niestroj, Lea Saller, Angela Schäfer, Tamara Schäfer, Janina Wayland und Melissa Wenzel leisteten durch ihr Engagement bei den Datenerhebungen wertvolle Beiträge zu dieser Arbeit. Ebenso danke ich meinen ehemaligen KollegInnen Dr. Melanie Sauerland, Dr. Tanja Stucke und Dr. Jürgen Gehrke, die mir nicht nur fachlich kompetente AnsprechpartnerInnen waren, sondern auch liebe FreundInnen. Meiner neuen Kollegin Dipl.-Psych. Kerstin Wilhelm danke ich für ihre Entlastung bei meinen universitären Aufgaben.

Mein besonderer Dank gilt all jenen, die mir emotional zur Seite standen. Nicht nur das Studium, sondern auch die vorliegende Arbeit wären ohne die zuverlässige und liebevolle Unterstützung meiner Mutter Uta Breuer nicht möglich gewesen. Ebenso möchte ich meinem Bruder Dr. Henning Breuer für seine Verlässlichkeit, seine Begeisterungsfähigkeit und viele kluge Worte danken. Nicht zuletzt verdanke ich ihm die Entscheidung für ein Studium der Psychologie, einer Wissenschaft, die mich stets fasziniert hat und faszinieren wird. Schließlich möchte ich mich bei Rachel Altmann für viele anregende und hilfreiche Gespräche bedanken.

Gießen, im Mai 2008 *Maike Breuer*

ZUSAMMENFASSUNG

Die Unterscheidung wahrer und erfundener Aussagen ist in vielen Lebensbereichen von Bedeutung und stellt zudem eine wesentliche Voraussetzung für eine effiziente und faire Rechtssprechung im Gerichtsverfahren dar. Die vorliegende Arbeit umfasst insgesamt fünf Studien, die sich auf die folgenden drei zentrale Fragestellungen beziehen: (1) Woran glauben Personen zu erkennen, dass jemand lügt bzw. wahr aussagt (Studien 1 und 2), (2) wodurch lassen sich erfundene und wahre Aussagen tatsächlich voneinander unterscheiden (Studien 3 und 4), und (3) lässt sich die Fähigkeit von Beurteilern den Wahrheitsstatus von Aussagen richtig einzuschätzen durch eine kurze Anleitung zu empirisch fundierten Aussagemerkmalen verbessern (Studie 5)?

Wiederholt wurde gezeigt, dass sowohl Laien als auch ExpertInnen kaum dazu in der Lage sind wahre und erfundene Aussagen zu unterscheiden. Ihre Urteilsfähigkeit liegt nach metaanalytischen Befunden nur geringfügig über dem Zufallsniveau. Eine mögliche Erklärung besteht darin, dass Alltagsvorstellungen über Lügenindikatoren nicht mit objektiven Indikatoren von Täuschung korrespondieren. Ebenso wäre es möglich, dass Personen die kontextuellen Besonderheiten von Aussagen nicht ausreichend berücksichtigen. Die ersten beiden Studien untersuchten diese Erklärungsansätze.

In der ersten Studie wurden die Alltagsvorstellungen von Laien bezüglich einer Vielzahl non- und paraverbaler Täuschungsindikatoren erfasst. Dabei wurden die Täuschungssituation und die Gelegenheit zur Vorbereitung als kontextuelle Determinanten manipuliert. Es zeigten sich nur geringe Unterschiede in den subjektiven Annahmen in Abhängigkeit vom variierten Aussagekontext, doch große Diskrepanzen zu objektiven Täuschungsindikatoren.

Im Vergleich zu nonverbalen Merkmalen erlauben es inhaltliche Aussagemerkmale besser zwischen wahren und erfundenen Aussagen zu unterscheiden. Daher wurden in der zweiten Studie Alltagsvorstellungen zu

inhaltlichen Glaubhaftigkeitsmerkmalen erfasst, den sogenannten Aberdeen Report Judgment Scales (ARJS, Sporer, 1996/1998/2004). Erneut wurden die Täuschungssituation und die Gelegenheit zur Vorbereitung über die Vorgabe fiktiver Szenarien variiert. Die meisten inhaltlichen Aussagemerkmale wurden gemäß ihrer objektiven Differenzierungskraft als Wahrheitsindikatoren aufgefasst. Während sich vereinzelt Unterschiede in Abhängigkeit von der Täuschungssituation zeigten, waren keine Effekte der Vorbereitung auf die subjektiven Täuschungsannahmen festzustellen.

Beide Studien verweisen auf eine Tendenz von Laien, die Validität verschiedener Indikatoren zu überschätzen. Zudem wurden bedeutsame Moderatoren der objektiven Differenzierungskraft kaum beachtet.

In der dritten und vierten Studie wurden theoretisch fundierte inhaltliche Glaubhaftigkeitsmerkmale hinsichtlich ihrer Inter-Rater-Reliabilität und Validität evaluiert. Dazu wurden wahre und erfundene freie Berichte und Interviews zu persönlich bedeutsamen Lebensereignissen anhand der ARJS (Studie 3) und anhand einer Kurzform derselben, der Aberdeen Report Judgment Scales--Short Training Version--German (ARJS-STV-G, Sporer & Masip, 2007) analysiert (Studie 4). Zudem wurden moderierende Effekte der Vorbereitung und der Valenz des geschilderten Ereignisses auf die Validität dieser sozial-kognitiven Glaubhaftigkeitsmerkmale überprüft.

In der dritten Studie wurden die transkribierten Aussagen von vier intensiv trainierten Raterinnen anhand der ARJS beurteilt. Für die meisten Skalen ergaben sich zufrieden stellende Inter-Rater-Reliabilitäten. Zudem ließen sich diese durch die Zusammenfassung mehrerer unabhängiger Beurteilungen substantiell verbessern. Erwartungsgemäß erzielten wahre Aussagen höhere ARJS-Beurteilungen als erfundene. Die Validität der ARJS war unabhängig von der Valenz der geschilderten Ereignissse und der Vorbereitung nachweisbar. Dies unterstützt die Zielsetzung eines breiten Anwendungsbereichs dieser Kriterien.

In der vierten Studie wurde dasselbe Stimulusmaterial von insgesamt acht Raterinnen anhand der ARJS-STV-G beurteilt. Durch die Zusammenfassung mehrerer unabhängiger Beurteilungen wurden auch für die meisten Merkmale der Kurzform zufrieden stellende Inter-Rater-Reliabilitäten erzielt. Erwartungsgemäß wiesen wahre Aussagen im Vergleich zu erfundenen eine höhere inhaltliche Qualität auf. Es war kein moderierender Effekt der Vorbereitung auf die Validität der ARJS-STV-G festzustellen. Insgesamt verwiesen die Untersuchungsbefunde darauf, dass es nützlich sein könnte, Laien anhand der ökonomischen Kurzform über inhaltliche Glaubhaftigkeitsmerkmale zu informieren.

Die fünfte Studie untersuchte die Effekte einer ARJS-STV-G-Anleitung auf die Urteilsgüte. Dazu wurden die subjektiven Glaubhaftigkeitsurteile von Beurteilerinnen analysiert, die jeweils einen Teil des Stimulusmaterials naiv und unter Anleitung der ARJS-STV-G beurteilten. Für die naiven Urteile zeigte sich eine rein zufällige Urteilsrichtigkeit. Hingegen ergab sich für die ARJS-STV-G-angeleiteten Beurteilungen der Interviews eine überzufällige Urteilsrichtigkeit. Diese war auf eine verbesserte Einschätzung der wahren Aussagen zurückzuführen. Allerdings verstärkte sich auch die Tendenz, Aussagen als glaubhaft zu beurteilen. Die ARJS-STV-G wird daher insbesondere BeurteilerInnengruppen empfohlen, die einen Lügenbias aufweisen.

INHALTSVERZEICHNIS

STUDIE 3

INTER-RATER-RELIABILITÄT UND VALIDITÄT DER ABERDEEN REPORT JUDGMENT SCALES 93

STUDIE 4

INTER-RATER-RELIABILITÄT UND VALIDITÄT DER ABERDEEN REPORT JUDGMENT SCALES--SHORT TRAINING VERSION--GERMAN 208

EINFÜHRUNG

Jeder Richter[1] im deutschen Rechtssystem hat sich mit seinem Amtseintritt durch folgenden öffentlichen Eid der Wahrheitsfindung verpflichtet:

> "Ich schwöre, das Richteramt getreu dem Grundgesetz für die Bundesrepublik Deutschland und getreu dem Gesetz auszuüben, nach bestem Wissen und Gewissen ohne Ansehen der Person zu urteilen und nur der Wahrheit und Gerechtigkeit zu dienen […]." Deutsches Richtergesetz (DriG), 5. Abschnitt, §38 – Urteil vom 19. April 1972.

Die Wahrheit ist jedoch oft schwer zu finden. Dies gilt insbesondere für Fälle, in denen Aussage gegen Aussage steht und externe Beweismittel fehlen. Die rechtspsychologische Forschung hat vielfältige Phänomene aufgezeigt, die den Wahrheitsgehalt von Aussagen beeinträchtigen können. Dabei handelt es sich zum einen um unbeabsichtigte Irrtümer, die beispielsweise aufgrund regulärer Wahrnehmungs- und Gedächtnisprozesse entstehen oder auch durch suggestive Beeinflussungen von außen (z.B. Loftus, 1979; Yarmey, 1979). Zum anderen ist gerade im rechtspsychologischen Anwendungsfeld damit zu rechnen, dass Personen gezielt falsch aussagen, um unangenehme Konsequenzen zu vermeiden. Gegenstand der vorliegenden Arbeit sind solche intentionalen Falschaussagen, die im Folgenden als Lügen oder auch als erfundene Aussagen bezeichnet werden.

Die bisherige Forschung hat gezeigt, dass Personen kaum dazu in der Lage sind, wahre und erfundene Aussagen zu unterscheiden. Dies gilt sowohl für Laien (Bond & DePaulo, 2006) als auch für Personengruppen, die beruflich

---

[1] Zugunsten der Lesbarkeit werden im Folgenden männliche Bezeichnungen gewählt, um Personengruppen zu beschreiben. Dabei sind inhaltlich grundsätzlich auch Frauen mit eingeschlossen. An den nachfolgenden Untersuchungen waren teilweise ausschließlich Frauen beteiligt. Dies wird durch die Verwendung weiblicher Bezeichnungen indiziert.

vermutlich häufig mit Falschaussagen konfrontiert werden, wie Polizisten, Psychiater und Richter (Aamodt & Custer, 2006). Eine erfolgreiche Lügendetektion setzt voraus, dass tatsächlich Unterschiede zwischen wahren und erfundenen Aussagen bzw. in den damit einhergehenden Verhaltensweisen vorzufinden sind (objektive Indikatoren). Zum anderen erscheint es notwendig, dass die Annahmen von Personen über das Verhalten von wahr aussagenden bzw. lügenden Personen (subjektive Indikatoren) mit diesen objektiven Indikatoren korrespondieren. Fehlerhafte Annahmen im Sinne von Alltagsvorstellungen könnten demnach zur schlechten Urteilsfähigkeit bei der Entdeckung von Täuschung beitragen.

Die ersten beiden Untersuchungen zielten darauf ab, solche Alltagsvorstellungen bzw. subjektive Indikatoren von Täuschung näher zu untersuchen. Während in der ersten Studie non- und paraverbale Merkmale sowie allgemeine Gesamteindrücke von Interesse waren, wurden in der zweiten Studie subjektive Annahmen zu inhaltlichen Aussagemerkmalen untersucht. Im Gegensatz zu früheren Untersuchungen wurde die Abhängigkeit dieser Alltagstheorien von situativen Faktoren genauer untersucht.

Metaanalytisch wurden bislang nur geringe Zusammenhänge zwischen non- und paraverbalen Verhaltensweisen und Täuschung nachgewiesen (DePaulo, Lindsay, Malone, Muhlenbruck, Charlton & Cooper, 2003; Sporer & Schwandt, 2006, 2007). Dies steht im Kontrast zu der subjektiven Annahme verschiedener Berufsgruppen, dass Lügner nervöse Verhaltensweisen zeigen (z.B. Akehurst, Köhnken, Vrij & Bull, 1996; Strömwall & Granhag, 2003). Allerdings wurde oft nicht kontrolliert, auf welche Täuschungssituation diese subjektiven Annahmen bezogen sind. So wäre es denkbar, dass Studierende ihre Angaben auf Alltagslügen beziehen, während Polizisten an Vernehmungen von Tatverdächtigen denken. Für die objektive Differenzierungskraft non- und paraverbaler Merkmale scheint der Aussagekontext durchaus bedeutsam zu sein. So wurden unter anderem das Ausmaß an Motivation und Vorbereitung sowie der

Gegenstand der Lüge als Moderatoren objektiver Täuschungsindikatoren identifiziert (z.B. DePaulo et al., 2003).

Die erste Studie überprüfte, ob Personen ihre subjektiven Annahmen zu Täuschungskorrelaten nach den kontextuellen Besonderheiten von Aussagen ausrichten. Dazu wurde ein Fragebogen konzipiert, der eine Vielzahl der in der Literatur diskutierten non- und paraverbalen Indikatoren von Täuschung sowie auf den Gesamteindruck bezogene Merkmale umfasste. Die Versuchspersonen sollten angeben, ob sie diese Indikatoren eher mit wahren oder erfundenen Aussagen assoziierten. Ihre Angaben sollten sich dabei auf ein vorgegebenes fiktives Szenario beziehen. Durch diese Szenarien wurden die Gelegenheit zur Vorbereitung einer Falschaussage und der Gegenstand der Lüge systematisch variiert. Daraus resultierten sechs Untersuchungsgruppen, deren subjektive Annahmen einander vergleichend gegenübergestellt und mit dem Forschungsstand zu objektiven Indikatoren kontrastiert wurden.

Inhaltliche Aussagemerkmale scheinen besser dazu geeignet zu sein, zwischen wahren und erfundenen Aussagen zu unterscheiden als non- und paraverbale Verhaltensweisen (vgl. DePaulo et al., 2003). Im Rahmen der zweiten Untersuchung wurden subjektive Annahmen zu den inhaltlichen Aussagemerkmalen der Aberdeen Report Judgment Scales (ARJS; Sporer, 1996/1998/2004) erfasst. Dabei wurden erneut die Vorbereitung und der Gegenstand der Lüge über die Vorgabe fiktiver Szenarien variiert. Zudem wurde an die Überlegungen von Niehaus, Krause und Schmidke (2005) anknüpfend, anhand eines weiteren fiktiven Szenarios überprüft, ob es spezifische subjektive Annahmen für Falschaussagen bei Sexualdelikten gibt. Die subjektiven Annahmen der sieben Experimentalgruppen wurden zusätzlich mit denen einer Kontrollgruppe verglichen, die den Fragebogen ohne die Instruktion, an ein konkretes Szenario zu denken, bearbeitete.

Die beiden nachfolgenden Studien überprüften die objektive Differenzierungskraft sowie die Inter-Rater-Reliabilität bei der Beurteilung

inhaltlicher Glaubhaftigkeitsmerkmale. Im Rahmen der Criteria-Based Content Analysis (CBCA, Steller & Köhnken, 1989) wurden die bekanntesten inhaltlichen Aussagemerkmale vorgestellt. Allerdings wurden diese wiederholt wegen ihrer mangelnden theoretischen Fundierung kritisiert (z.B. Steller & Köhnken, 1989; Sporer, 1997a). Daher wurden aus dem Realitätsüberwachungsansatz (Johnson & Raye, 1981) andere inhaltliche Merkmale zur Unterscheidung wahrer und erfundener Aussagen abgeleitet (Sporer & Küpper, 1995; s. Masip, Sporer, Garrido & Herrero, 2005; Sporer, 2004). Sporer und seine Arbeitsgruppe wiederum arbeiteten Gemeinsamkeiten und Unterschiede zwischen den CBCA und RÜ-Merkmalen heraus (Sporer, 1997a, 1997b; Sporer & Bursch, 1996). Auf der Grundlage dieser Untersuchungen sowie von Theorien und Forschungsbefunden zur Attribution und zum autobiographischen Gedächtnis, stellte Sporer (1996/1989/2004) schließlich mit den ARJS 52 inhaltliche Glaubhaftigkeitsmerkmale vor.

Die dritte Studie zielte darauf ab, die Inter-Rater-Reliabilität und Validität der ARJS zu überprüfen. Dazu wurden vier Raterinnen, die sich bereits zuvor mit der Täuschungsliteratur beschäftigt hatten, intensiv in der Durchführung von ARJS-Analysen geschult. Sie beurteilten umfangreiche transkribierte Aussagen, die entweder wahr oder erfunden waren. Da die ARJS einen breiten Geltungsbereich beanspruchen, wurden Aussagen zu verschiedenen persönlich bedeutsamen Lebensereignissen als Stimulusmaterial verwendet. Die Befunde zur Validität ergänzend wurden moderierende Effekte der Aussageform, der Gelegenheit zur Vorbereitung und der Valenz der geschilderten Ereignisse analysiert.

Eine vollständige ARJS-Analyse erfordert fundierte Kenntnisse der relevanten Literatur und ist auch in der Durchführung äußerst aufwändig. Im Rahmen der vierten Studie wurde daher überprüft, ob auch Beurteilerinnen ohne Vorkenntnisse dazu in der Lage sind, die Glaubhaftigkeitsmerkmale reliabel und valide anzuwenden. Dazu wurde dasselbe Stimulusmaterial, das bereits in der dritten Studie verwendet wurde, acht weiteren Raterinnen zur Beurteilung

vorgelegt. Diese schätzten zunächt die Glaubhaftigkeit für einen Teil der Aussagen naiv ein. Danach wurden sie über 17 Glaubhaftigkeitsmerkmale informiert, die in einer Kurzform der ARJS enthalten sind, der Aberdeen Report Judgment Scales--Short Training Version--German. Die Beurteilungen der ARJS-STV-G-Merkmale wurden auf ihre Inter-Rater-Reliabilität und Validität hin überprüft. Dabei wurden erneut auch die Effekte der Aussageform, der Gelegenheit zur Vorbereitung und der Valenz der geschilderten Ereignisse auf die Einschätzung der Glaubhaftigkeitsmerkmale betrachtet.

Schließlich blieb ungeklärt, ob die Beurteilerinnen auch bei der Glaubhaftigkeitsbeurteilung von ihrem empirisch fundierten Wissen profitierten. Die subjektiven Glaubhaftigkeitsurteile wurden daher verwendet, um in einer fünften Studie die Urteilsgüte, -neigung und –sicherheit naiver (d.h. ohne Anleitung) und ARJS-STV-G-geschulter Beurteilerinnen zu vergleichen.

Literatur

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar: A meta-analysis of individual differences in detecting deception. Forensic Examiner, 15, 6-11.

Akehurst, L., Köhnken, G., Vrij, A., & Bull, R. (1996). Lay persons' and police officers' beliefs regarding deceptive behaviour. Applied Cognitive Psychology, 10, 461-471.

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. Personality and Social Psychology Review, 10, 214-234.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. Psychological Bulletin, 129, 74-112.

Deutsches Richtergesetz DriG, 5. Abschnitt, §38 – Urteil vom 19. April 1972. Aufgerufen im April 2007, von: http://bundesrecht.juris.de/drig/BJNR016650961.html.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. Psychological Review, 88, 67-85.

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

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. Psychology, Crime, and Law, 11, 99-122.

Niehaus, S., Krause, A., & Schmidke, J. (2005). Täuschungsstrategien bei der Schilderung von Sexualstraftaten. Zeitschrift für Sozialpsychologie, 36, 175-187.

Sporer, S. L. (1996/1998/2004). The Aberdeen Report Judgment Scales (ARJS). Definitions and answer sheets. Unpublished Questionnaire. University of Aberdeen, Scotland; University of Giessen, Germany.

Sporer, S. L. (1997a). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. Applied Cognitive Psychology, 11, 373-397.

Sporer, S. L. (1997b). Realitätsüberwachungskriterien und forensische Glaubwürdigkeitskriterien im Vergleich. In L. Greuel, T. Fabian, & M. Stadler (Eds.), Psychologie der Zeugenaussage (S. 71-85). München: Psychologie Verlags Union.

Sporer, S. L. (2004). Reality monitoring and detection of deception. In P. A. Granhag & L. Strömwall (Eds.), Deception detection in forensic contexts (pp. 64-102). Cambridge: University Press.

Sporer, S. L., & Bursch, S. E. (1996, April). Detection of deception by verbal means: Before and after training. Paper presented at the 38. Tagung experimentell arbeitender Psychologen, Eichstaett, Germany.

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

Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. Applied Cognitive Psychology, 20, 421-446.

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. Psychology, Public Policy, and Law, 13, 1-34.

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

Strömwall, L. A., & Granhag, P. A. (2003). How to detect deception? Arresting the beliefs of police officers, prosecutors and judges. Psychology, Crime, and Law, 9, 19-36.

Yarmey, A. D. (1979). The psychology of eyewitness testimony. New York: Free Press.

STUDIE 1

Subjektive Indikatoren von Täuschung: Die Bedeutung der Täuschungssituation und Gelegenheit zur Vorbereitung

Für gerichtliche Entscheidungen ist es von zentraler Relevanz, wahre von falschen Aussagen zu unterscheiden. Dies gilt insbesondere für Fälle in denen wenig eindeutige Beweismittel vorliegen (z.B. bei Sexualdelikten). Es ist jedoch davon auszugehen, dass Richter in den meisten Prozessen die Glaubhaftigkeit von Aussagen ohne die Hilfestellung forensischer Sachverständiger bewerten (Volbert & Busse, 1995). Nicht nur Laien, sondern auch Berufsgruppen, die häufig Aussagen hinsichtlich ihrer Glaubhaftigkeit beurteilen, erreichen oft nur Trefferquoten im Bereich der Zufallswahrscheinlichkeit. Dies berichten beispielsweise Ekman und O`Sullivan (1991) für Richter, Psychiater und Polizisten (vgl. auch Köhnken, 1990; Strömwall & Granhag, 2003; Vrij, 2000).

Um dieses Phänomen zu verstehen, sind die Ergebnisse zweier Forschungsrichtungen relevant. Zum einen stellt sich die Frage nach objektiven Täuschungsindikatoren, also ob es überhaupt Unterschiede im Verhalten von Lügnern und Personen, die wahre Aussagen formulieren, gibt. Zum anderen sollten für eine erfolgreiche Lügendetektion die objektiven mit den subjektiven Täuschungsindikatoren übereinstimmen, also mit den Annahmen von Personen darüber, welche Verhaltensweisen kritisch für die Unterscheidung wahrer und falscher Aussagen sind (vgl. das Brunswiksche Linsenmodell: Fiedler, 1989a; Köhnken, 1990; Reinhard, Burghardt, Sporer & Bursch, 2002; Sporer, 1997). Wenn Personen auf die falschen Verhaltensweisen achten, um den Wahrheitsgehalt von Aussagen zu beurteilen, können sie bei der Lügendetektion auch nicht erfolgreich sein.

Die untersuchten objektiven und subjektiven Täuschungsindikatoren lassen sich in drei Kategorien unterteilen. Nonverbale Indikatoren umfassen direkt beobachtbare Verhaltensweisen wie Blickkontakt und Bewegungsverhalten.

Paraverbale Indikatoren beziehen sich auf Verhaltensweisen, die meist nicht vollkommen unabhängig vom Inhalt zu erfassen sind, wie die Stimmlage, Antwortlatenzen und Redepausen. Inhaltliche Indikatoren, wie die Widerspruchsfreiheit einer Aussage, setzen hingegen eine inhaltliche Aussageanalyse voraus. Die vorliegende Untersuchung bezieht sich auf nonverbale und paraverbale Täuschungsindikatoren.

Objektive Täuschungsindikatoren

Zuckerman, DePaulo und Rosenthal (1981) haben vier theoretische Ansätze formuliert, die objektive Verhaltensunterschiede in Abhängigkeit vom Wahrheitsstatus einer Aussage erwarten lassen, den Erregungs-, Kontroll-, Kognitiven und Affektiven Ansatz. Diese verschiedenen Ansätze basieren auf der Annahme, dass es kein spezifisches Lügenverhalten gibt, also kein Verhalten, das immer auftritt, wenn jemand lügt, aber nie beobachtbar wäre, wenn jemand die Wahrheit sagt. Vielmehr wird spezifiziert, welche Gedanken, Gefühle und psychologischen Prozesse mit Lügen einhergehen, und in welchen Verhaltensweisen sie sich manifestieren.

So ist nach dem Erregungsansatz die autonome Reaktion bei Lügnern erhöht, was eine Zunahme nervöser Verhaltensweisen erwarten lässt. Der Kontrollansatz geht hingegen davon aus, dass Lügner aktiv versuchen diejenigen Verhaltensweisen zu kontrollieren, von denen sie annehmen, dass Lügenentdecker diese als Indikatoren von Täuschung annehmen. Daher wird beispielsweise eine Abnahme von Bewegungsverhalten postuliert. Der Kognitive Ansatz verweist auf die erhöhte Belastung des Arbeitsgedächtnisses bei der Formulierung von komplexen erfundenen im Gegensatz zu wahren Aussagen (Sporer & Schwandt, 2006, 2007) und argumentiert, dass es infolgedessen zu verzögerten Antworten und vermehrten Pausen kommen kann. Der Affektive Ansatz expliziert, dass Lügen mit Schuldgefühlen, Angst oder auch Freude über die Herausforderung erfolgreich zu täuschen (duping delight) einhergehen (Ekman, 1992), was jeweils zu unterschiedlichen Verhaltensänderungen führt. Diese

Ansätze sind nicht als konfligierende Theorien zu betrachten. Vielmehr ergänzen sie sich dabei, bisherige Befunde zu objektiven Täuschungsindikatoren zu erklären. Kürzlich wurden diese klassischen Ansätze von DePaulo et al. (2003) um die sogenannte Selbstdarstellungstheorie ergänzt, die sich hauptsächlich auf Lügen im Alltagskontext bezieht. Nach diesem Ansatz sind Personen im Allgemeinen bemüht, sich in einem positiven Licht darzustellen und den Eindruck, den sie auf andere machen, durch strategische verbale und nonverbale Kommunikation zu beeinflussen (vgl. hierzu Fiedler, 1989).

Es ist jedoch festzuhalten, dass die tatsächlichen Unterschiede zwischen wahren und falschen Aussagen meist nur gering ausfallen. DePaulo et al. (2003) untersuchten insgesamt 88 Hinweise aus 120 unabhängigen Studien daraufhin, ob sie zwischen wahren und gelogenen Aussagen differenzieren. Der Median der ermittelten Effektstärken lag bei d = .10 (vgl. auch Tabelle 1), wobei nach Cohen (1988) d = .20 als geringer, d = .50 als mittlerer und d = .80 als starker Effekt zu interpretieren ist.

Moderatoren bei objektiven Indikatoren von Täuschung

Ein Grund für die geringe Differenzierungskraft nonverbaler Hinweise ist darin zu sehen, dass es eine Vielfalt von Begebenheiten gibt, in denen auf unterschiedliche Weise gelogen wird. Die Motivation zu überzeugen, die Gelegenheit zur Vorbereitung und der Aussageinhalt wurden unter anderem als bedeutsame Moderatoren identifiziert. So fanden DePaulo et al. (2003) generell stärkere Effekte bei Studien mit motivationalen Anreizen. Nur bei hoher Motivation erschienen Lügner angespannter und sprachen in einer höheren Stimmlage als Personen, die wahr aussagten. Wenn die Motivation als gering einzuschätzen war, zeigten sich diesbezüglich keine Unterschiede. Im Vergleich zu Personen, die wahre Aussagen formulierten, zeigten Lügner nur in Bedingungen mit hoher Motivation weniger Blickkontakt, und ihre Aussagen beinhalteten weniger gefüllte Pausen. Bei niedriger Motivation zeigte sich hingegen ein umgekehrtes Muster.

Tabelle 1
Effektstärken subjektiver Täuschungsindikatoren als Funktion der Vorbereitungszeit und Situation sowie Vergleich zu objektiven Täuschungsindikatoren

	Subjektiv (vorliegende Studie)						Objektiv (DePaulo et al., 2003)	
	Vorbereitung		Situation					
Indikatoren	Ohne	Mit	Alltag	Affäre	Verbrechen	d_{Subj}	d_{Obj}	k
Sprachverhalten								
Ungefüllte Sprechpausen	0.05	-0.14	-0.19	0.14	-0.06	-0.04	0.01	15
Gefüllte Sprechpausen	0.88	0.77	0.89	0.95	0.64	**0.82**	0.00	16
Stottern	0.67	0.54	0.57	0.66	0.58	**0.60**	0.22	1
Räuspern, Hüsteln	0.52	0.72	0.58	0.49	0.79	**0.63**	0.00	17
Versprecher	0.98	0.72	0.84	1.11	0.64	**0.84**	0.00	17
Gramm. Fehler, Unvollständige Sätze [a]	0.49	0.19	0.30	0.39	0.32	**0.34**	0.00	17
Stereotype Vorstellungen	0.75	0.62	0.69	0.70	0.64	**0.68**	--	--
Floskeln, Füllwörter	0.64	0.54	0.65	0.59	0.58	**0.59**	0.20	4
Antwortverzögerungen [b]	0.87	0.66	0.77_{xy}	1.23_{x}	0.46_{y}	**0.76**	0.02	32
Hektische Redeweise	1.02	0.82	0.94	1.12	0.73	**0.92**	0.07	23
Stockende Redeweise [a]	0.57	0.27	0.45_{xy}	0.60_{x}	0.22_{y}	**0.42**	0.17	8
Hohe Stimme	0.30	0.12	0.29	0.18	0.14	0.21	0.21	12
Monotoner Tonfall [a]	-0.74	-0.42	-0.48	-0.70	-0.55	**-0.57**	-0.42	1
Zittrige Stimme	0.54	0.31	0.34	0.50	0.42	**0.42**	**0.30**	10
Leise Stimme	0.07	-0.14	0.03	-0.15	-0.19	-0.10	-0.05	5

Tabelle 1 (Fortsetzung)

	Subjektiv						Objektiv	
	Vorbereitung		Situation					
Indikatoren	Ohne	Mit	Alltag	Affäre	Verbrechen	$\underline{d}_{Subj}$	$\underline{d}_{Obj}$	$\underline{k}$
Angespannte Stimme	0.88	1.27	1.13	1.01	1.00	**1.05**	0.26	10
Unfreundlicher Tonfall	0.24	0.28	0.40	0.13	0.27	**0.26**	**-0.11**	4
Umfang des Wortschatzes	0.03	0.06	0.12	-0.05	0.06	0.04	-0.07	6
Ausführlichkeit der Antworten [a]	0.01	-0.28	-0.15	0.01	-0.28	-0.13	-0.03	49
Hoher Redeanteil [ab]	0.20	-0.53	-0.56_{x}	-0.16_{y}	-0.34_{xy}	**-0.35**	**-0.35**	4
Längere Sätze [a]	0.21	-0.13	-0.05	0.14	0.04	0.05	-0.07	6
Gramm. komplizierte Sätze	0.39	0.25	0.29	0.28	0.39	**0.32**	-0.07	6
Verhalten im Gesichtsbereich								
Dauer fehlenden Blickkontaktes	0.72	1.20	1.04	0.94	0.78	**0.92**	0.01	32
Häufigkeit der Blickabwendung	0.94	0.96	1.02	1.12	0.77	**0.95**	0.03	6
Umherirrender Blick	0.67	0.78	0.71	0.78	0.66	**0.72**	-0.07	11
Erweiterte Pupillen	0.22	0.30	0.41	0.09	0.27	**0.26**	0.39	4
Lidschlag	0.54	0.49	0.45	0.63	0.48	**0.51**	0.07	17
Zuckungen im Gesicht	0.46	0.34	0.44	0.30	0.48	**0.40**	**0.29**	1
Augenbrauen zusammenziehen	0.44	0.23	0.40	0.36	0.24	**0.33**	0.04	5
Bewegungen der Nase	0.05	0.01	-0.03	0.03	0.08	0.03	--	--
Kinn anheben/vorstrecken	0.00	-0.01	0.04	0.01	-0.06	0.00	0.25	4

Tabelle 1 (Fortsetzung)

	Subjektiv						Objektiv	
	Vorbereitung		Situation					
Indikatoren	Ohne	Mit	Alltag	Affäre	Verbrechen	d_{Subj}	d_{Obj}	k
Lachen [b]	0.06	-0.08	0.21_x	0.11_x	-0.29_y	-0.01	0.00	27
Grinsen [b]	0.08	0.10	0.26_x	0.24_x	-0.18_y	0.09	0.02	4
Lächeln [b]	0.05	-0.03	0.21_x	0.13_x	-0.31_y	0.01	-0.70	2
Gekünsteltes Lächeln	0.99	1.06	1.33	1.01	0.82	**1.02**	**0.31**	2
Lippen zusammenpressen	0.70	0.50	0.43	0.67	0.74	**0.61**	**0.16**	4
Schlucken	0.72	0.95	0.72	0.87	0.88	**0.82**	--	--
Erröten im Gesicht	1.14	0.95	0.97	1.27	0.94	**1.05**	--	--
Blass werden im Gesicht [b]	0.16	-0.01	-0.16_x	0.18_{xy}	0.23_y	0.08	--	--
Ausdrucksstarke Mimik	0.00	0.10	0.05	-0.01	0.12	0.05	0.12	3
Verkrampfter Gesichtsausdruck	0.75	0.58	0.73	0.50	0.79	**0.66**	--	--
Unfreundl. Gesichtsausdruck	0.22	0.25	0.40	0.18	0.12	**0.23**	**0.12**	13
Kopfnicken [b]	0.31	0.36	0.26_{xy}	0.17_x	0.55_y	**0.33**	0.01	16
Kopfschütteln [b]	0.28	0.11	0.05_x	0.46_x	0.12_x	0.20	-0.12	5
Nervöser Gesamteindruck im Kopfbereich	1.30	1.14	1.22	1.25	1.20	**1.21**	0.29	1
Verhalten im Körperbereich								
Veränderung der Körperhaltung	0.82	0.90	0.69	1.14	0.82	**0.86**	0.05	29
Zuwendung des Körpers	0.35	0.37	0.39	0.30	0.39	**0.36**	-0.20	2
Verschlossene Körperhaltung	0.61	0.71	0.68	0.83	0.50	**0.66**	--	--

Tabelle 1 (Fortsetzung)

	Subjektiv						Objektiv	
	Vorbereitung		Situation					
Indikatoren	Ohne	Mit	Alltag	Affäre	Verbrechen	d_{Subj}	d_{Obj}	k
Verkrampfte Körperhaltung	1.06	0.92	0.96	0.94	1.07	**0.99**	0.02	13
Zittern	0.60	0.38	0.32	0.50	0.66	**0.49**	0.11	4
Achselzucken	0.13	0.33	0.22	0.23	0.23	0.23	0.04	6
Armbewegungen	0.75	0.73	0.90	0.67	0.67	**0.74**	-0.17	3
Erläuternde Gesten	1.09	0.93	1.27	0.84	0.98	**1.02**	**-0.14**	16
Bewegungen der Hände	1.10	1.18	1.50	1.00	1.02	**1.14**	0.00	29
Manipulationen am Körper	1.33	1.49	1.68	1.38	1.24	**1.42**	-0.01	18
Hantieren mit Gegenständen	1.30	1.38	1.40	1.44	1.17	**1.33**	-0.12	5
Bewegungen der Beine	0.76	0.89	0.70	0.90	0.89	**0.83**	-0.09	28
Bewegungen der Füße	0.72	0.97	0.80	0.81	0.90	**0.84**	-0.09	28
Nervöser Gesamteindruck im Körperbereich	1.35	1.31	1.28	1.42	1.20	**1.32**	--	--
Gesamteindruck								
Widersprüche zwischen Ausdruck im Gesicht und Sprechverhalten	0.97	0.80	0.69	0.79	1.28	**0.88**	**0.34**	7
Widersprüche zwischen Inhalt der Aussage und Körpersprache	1.05	1.09	1.11	0.95	1.18	**1.07**	**0.34**	7
Ruhig/nervös [a]	1.30	0.74	0.91	1.02	1.01	**0.98**	**0.27**	16
Angespannt/entspannt	-0.45	-0.35	-0.36	-0.26	-0.61	**-0.40**	**0.27**	16
Negativ/positiv	0.27	-0.10	-0.11	-0.14	-0.32	-0.18	0.66	3

Tabelle 1 (Fortsetzung)

	Subjektiv						Objektiv	
	Vorbereitung		Situation					
Indikatoren	Ohne	Mit	Alltag	Affäre	Verbrechen	d_{Subj}	d_{Obj}	k
Unfreundlich/freundlich	0.24	0.29	0.33	0.28	0.20	**0.28**	0.16	6
Distanziert/engagiert	0.13	0.21	0.19	0.24	0.09	0.17	0.08	6
Abstrakt/konkrete Äußerungen	0.15	-0.01	-0.11	-0.07	-0.06	-0.08	**-0.30**	24
Vage/lebendig	0.13	-0.09	0.05	-0.29	-0.12	-0.11	**0.30**	10
Unklar/klar	0.15	-0.02	-0.03	-0.22	-0.02	-0.09	**-0.55**	7
Zäh/flüssig	0.29	0.11	-0.09	-0.31	0.09	-0.09	0.17	8
Person ist abgewendet/zugewendet	-0.35	-0.17	-0.28	-0.33	-0.19	**-0.27**	0.07	11
Denkt nicht nach/hart nach [b]	0.64	0.67	0.45_{x}	0.92_{y}	0.62_{xy}	**0.65**	0.61	1
Unangenehm/angenehm [ab]	-1.34	-0.92	-1.00_{x}	-1.07_{xy}	-1.26_{y}	**-1.11**	--	--

Anm. Für objektive Täuschungsindikatoren k = Anzahl unabhängiger Einzelstudien nach DePaulo et al. (2003); d_{Obj} = Effektstärke der objektiven Differenzierungskraft nach DePaulo et al. (2003).

[a] = Signifikanter Haupteffekt für Vorbereitungszeit mit df = 1,228; [b] = Signifikanter Haupteffekt für Situation mit df = 2,228; Zellenmittelwerte mit unterschiedlichen Subskripta x, y und z unterscheiden sich signifikant ($p < .017$).

Welche Bedeutung der Vorbereitung einer Aussage zukommt, wurde in den Metaanalysen von Sporer und Schwandt (2006, 2007) untersucht. Bei kürzerer Vorbereitungszeit wurden eine höhere Stimmlage und längere Antwortlatenzen bei Lügnern beobachtet. Um den Gegenstand der Lüge als Moderator zu berücksichtigen, differenzierten DePaulo et al. (2003) zwischen Lügen über eigene Verfehlungen, die beispielsweise über unerlaubtes Abschreiben operationalisiert wurden, einerseits und Lügen über persönliche Meinungen und Erfahrungen andererseits. Die Effekte fielen bei Aussagen über eigene Verfehlungen stärker aus.

Zusammenfassend gibt es zwar objektive Unterschiede im Verhalten von Lügnern und Personen, die wahre Aussagen formulieren, sie fallen jedoch so gering aus, dass sie praktisch unbedeutsam sind, solange man nicht wichtige Moderatorvariablen berücksichtigt.

Subjektive Täuschungsindikatoren

Subjektive Täuschungsindikatoren sind Verhaltensweisen, von denen Personen annehmen, dass sie mit Lügen einhergehen. Diese werden in der Regel über drei Paradigmen erfasst: Fragebogenuntersuchungen, Beurteilungsstudien und Beurteilungsstudien, bei denen die Versuchspersonen ihr Urteil begründen. Bei Fragebogenuntersuchungen werden Verhaltensweisen aufgeführt, und die Versuchspersonen sollen angeben, ob sie diese als Lügenindikatoren einschätzen. Es werden also subjektive Annahmen darüber erfasst, welche der vorgegebenen Verhaltensweisen mit Lügen assoziiert werden. Hingegen bleibt offen, ob diese Merkmale auch tatsächlich genutzt werden, um den Wahrheitsgehalt einer Aussage zu beurteilen. Dies zu erfassen ist im Rahmen von Beurteilungsstudien möglich, bei denen Versuchspersonen einschätzen, ob eine auf Video dargebotene Aussage wahr oder erlogen ist. Das Urteil wird dann im Sinne einer Brunswikschen Linsenanalyse (Fiedler, 1989; Köhnken, 1990; Sporer, 1997) den tatsächlich beobachtbaren Verhaltensweisen gegenübergestellt. Daraus lässt sich ableiten, welche Verhaltensweisen als

Täuschungsindikatoren wahrgenommen werden. Unklar bleibt jedoch, ob sich die Personen darüber bewusst sind, welche Verhaltensweisen sie für die Beurteilung genutzt haben (vgl. Nisbett & Wilson, 1977). Nach Zuckerman, Koestner und Driver (1981) korrelieren die Ergebnisse von Fragebogen- und Beurteilungsstudien jedoch sehr hoch. Demnach ist davon auszugehen, dass Personen auch tatsächlich die Hinweise zur Beurteilung des Wahrheitsstatus einer Aussage nutzen, die sie zu nutzen glauben. Zuweilen wird den Versuchpersonen nach einer Beurteilung auch Gelegenheit gegeben, anzugeben, wie sie zu ihrem Urteil gekommen sind. Problematisch hierbei ist jedoch, dass die nachträglich formulierten Gründe nicht notwendigerweise entscheidungsleitend gewesen sein müssen (vgl. Reinhard et al., 2002).

Die Forschungsbefunde zu subjektiven Täuschungsindikatoren zeigen, dass sich Personen über verschiedene Berufsgruppen hinweg relativ einig darin sind, welche Verhaltensweisen sie mit Täuschung assoziieren (Akehurst, Köhnken, Vrij & Bull, 1996; Hocking & Leathers, 1980; Köhnken, 1988; Zuckerman et al., 1981; Strömwall & Granhag, 2003). Sie teilen jedoch häufig falsche Annahmen darüber, wie sich Lügner verhalten. Gegenüberstellungen von Köhnken (1990) und Vrij (2000) lassen verschiedene Diskrepanzen zwischen subjektiven und objektiven Täuschungsindikatoren erkennen. Viele subjektive Annahmen über Täuschungsindikatoren scheinen einem Zappelphilipp-Stereotyp zu entsprechen. So wird beispielsweise erwartet, dass Lügner häufiger Illustratoren verwenden, Selbst-Manipulationen, Körper-, Hand- und Finger-, Bein- und Fußbewegungen zeigen und ihre Sitzhaltung verändern, als Personen, die wahr aussagen. Tatsächlich diskriminieren diese Verhaltensweisen jedoch entweder nicht zwischen den Gruppen oder nehmen sogar bei falschen Aussagen ab. Zudem achten Personen zu sehr auf leicht kontrollierbare Regionen, vor allem auf das Gesicht (Vrij, 2000). Es lässt sich somit in Übereinstimmung mit Köhnken (1990), Reinhard et al. (2002), Vrij (2000) und Zuckerman et al. (1981) festhalten, dass Personen ihre Glaubhaftigkeitsbeurteilung stark nach ihren stereotypen

Vorstellungen über typische Begleiterscheinungen von Lügen richten, diese Stereotype jedoch kaum mit den tatsächlichen Begleiterscheinungen von Lügen zusammenhängen.

Ziele und Hypothesen

Bei der Suche nach objektiven Täuschungsindikatoren hat es sich als hilfreich erwiesen, Moderatoren zu berücksichtigen. Es bleibt jedoch ungeklärt, welche Bedeutung Moderatoren für subjektive Annahmen über Täuschungsindikatoren zukommt.

Um diese Fragestellung zu untersuchen, wird in der vorliegenden Untersuchung ein Fragebogenparadigma genutzt. Dabei geben die Versuchpersonen an, ob vorgegebene Verhaltensweisen ihres Erachtens bei erlogenen im Vergleich zu wahren Aussagen zu- bzw. abnehmen. In fast allen bisherigen Untersuchungen wurden dabei weder der Inhalt noch die genauen Umstände der Aussage spezifiziert. Daher ist nicht auszuschließen, dass sich die subjektiven Täuschungsannahmen der befragten Personen auf völlig unterschiedliche Begebenheiten beziehen. Möglicherweise dachten die einen bei der Beantwortung des Fragebogens an eine Begebenheit, in der sie selbst eine triviale Lüge formuliert haben, andere daran wie ihr Partner sie belogen hat, um eine Affäre zu verheimlichen, und wieder andere an ein strafrechtlich relevantes Ereignis, von dem sie kurz zuvor durch die Medien erfahren haben.

Bei Studien, in denen unterschiedliche (Berufs-)gruppen miteinander verglichen wurden (z.B. Akehurst et al., 1996; Strömwall & Granhag, 2003), ist es wahrscheinlich, dass die jeweiligen Gruppen an ein für ihren Berufsalltag typisches Ereignis gedacht haben, d.h. Polizisten an ein Verhör, Richter an eine Vernehmung im Gerichtssaal und Studierende vermutlich eher an eine Alltagslüge. Die subjektiven Annahmen einer bestimmten Gruppe von Befragten wären dann mit dem vorgestellten Täuschungsereignis konfundiert. Daher wurden die Versuchspersonen in der vorliegenden Studie explizit instruiert, bei der

Beurteilung von Lügenindikatoren an ein vorgegebenes fiktives Szenario zu denken.

Ein entsprechendes Vorgehen wurde nach unserem Kenntnisstand bisher nur in einer Studie realisiert. Taylor und Vrij (2000) untersuchten subjektive Annahmen in Abhängigkeit von der Motivation der lügenden Person und der Komplexität der Aussage. Um die Komplexität der Aussage zu variieren, gaben sie fiktive Szenarien vor, bei denen entweder nur die Aussage des Lügners (geringe Komplexität) oder weitere Beweismittel, beispielsweise die konfligierenden Aussagen von Augenzeugen (hohe Komplexität), verfügbar waren. Durch diese Art der Operationalisierung bleibt jedoch unklar, ob Unterschiede in den subjektiven Annahmen wirklich auf die Komplexität der Aussage zurückzuführen sind oder vielmehr darauf, dass die von den Lügnern vorgebrachten Tatsachen widerlegbar waren. Daher wurde in der vorliegenden Untersuchung die kognitive Komplexität über das Ausmaß der vermeintlichen Vorbereitungszeit variiert, ohne die Beweisgrundlage zu verändern. Zudem beschrieben Taylor und Vrij (2000) die lügende Person durchgängig als männlich und erfassten nur eine begrenzte Anzahl von Verhaltensweisen. Im Gegensatz dazu wurden in der vorliegenden Untersuchung das Geschlecht der lügenden Person ausbalanciert, einzelne Verhaltensweisen ergänzt bzw. spezifiziert und zusätzlich subjektive Annahmen zum Gesamteindruck erfasst.

Die in der vorliegenden Untersuchung vorgegebenen fiktiven Szenarien unterschieden sich durch den Gegenstand der Lüge und die Gelegenheit zur Vorbereitung. Diese beiden Aspekte wurden orthogonal manipuliert, da sie sich bereits bei der Erforschung objektiver Verhaltensunterschiede als relevante Moderatoren erwiesen haben. Es wird erwartet, dass die subjektiven Annahmen zu Täuschungsindikatoren bisherigen Forschungsbefunden entsprechen, jedoch in Abhängigkeit vom vorgegebenen Szenario unterschiedlich stark ausgeprägt sind. Da viele Personen vermutlich im Alltag schon einmal eine Lüge erfinden mussten, wird angenommen, dass es ihnen bewusst ist, dass es schwieriger ist

spontan zu lügen, als nach ausreichender Vorbereitung. Daher sollten subjektive Annahmen zu Lügenindikatoren in den Bedingungen ohne Vorbereitungszeit deutlicher ausfallen als in Bedingungen mit Vorbereitungszeit. Zudem wird davon ausgegangen, dass die vorgegebenen Szenarien je nach Gegenstand der Lüge als unterschiedlich bedeutsam wahrgenommen werden. Es wird postuliert, dass es von der Alltagssituation mit geringen Konsequenzen über die Alltagssituation mit hohen Konsequenzen bis hin zur strafrechtlich relevanten Situation, zunehmend riskanter ist zu lügen. Daher sollte die Motivation erfolgreich zu täuschen zunehmen. Infolgedessen sollten unsere Versuchspersonen vor allem in der strafrechtlich relevanten Situation viele der vorgegebenen Verhaltensweisen als subjektiv relevante Lügenindikatoren einstufen. Hingegen sollten die Versuchspersonen annehmen, dass diese Verhaltensweisen in Alltagssituationen weniger stark ausgeprägt sind.

Methode

Zur Erfassung der subjektiven Annahmen zu Täuschungsindikatoren wurde 240 Versuchspersonen ein Fragebogen vorgelegt. Die Stichprobe bestand überwiegend aus Studierenden, fünf Personen waren in akademischen, vier in nicht-akademischen Bereichen tätig und 19 Personen machten keine beruflichen Angaben. Der Fragebogen umfasste 59 nonverbale und paraverbale Verhaltensweisen, die den drei Kategorien „Sprechverhalten“ (22 Items), „Verhalten im Gesichtsbereich“ (23 Items) und „Verhalten im Körperbereich“ (14 Items) zugeordnet waren. Weiterhin wurden über 14 Items subjektive Annahmen zum „Gesamteindruck“ erfasst. Die Versuchpersonen sollten für jedes Item auf einer 7-stufigen relativen Antwortskala von –3 bis +3 angeben, ob es seltener oder häufiger in erlogenen als in wahren Aussagen vorzufinden sei. Negative Werte indizierten eine Abnahme und positive Werte eine Zunahme des Verhaltens beim Lügen. Der Wert „0“ sollte angekreuzt werden, wenn die Versuchpersonen erwarteten, dass die Verhaltensweise weder zu- noch abnehmen würde.

Die Untersuchungsbedingungen unterschieden sich in der Vorgabe fiktiver Szenarien, auf die sich die Angaben zu subjektiven Täuschungsannahmen beziehen sollten. Dabei wurden die Täuschungssituation und die Gelegenheit zur Vorbereitung systematisch variiert. Um mögliche geschlechtsspezifische Einflüsse auszubalancieren, wurden gleich viele Männer und Frauen befragt sowie das Geschlecht der fiktiven Stimulusperson variiert.

In der Alltagssituation mit geringen Konsequenzen formuliert Frau/ Herr K. eine Lüge, um sich für eine Verspätung zu rechtfertigen (Verspätungssituation). Eine Alltagssituation mit hohen Konsequenzen wurde dadurch operationalisiert, dass Frau/ Herr K. ihren Partner/ seine Partnerin belügt, um eine außereheliche Affäre zu verheimlichen (Affärensituation). In der strafrechtlich relevanten Situation erschlägt Frau/ Herr K. ihren Partner/ seine Partnerin im Affekt bei einem Streit um einen Erbnachlass und sagt gegenüber der Polizei falsch aus, um sich nicht dafür verantworten zu müssen (Verbrechenssituation). Um das Ausmaß an Vorbereitungszeit zu variieren, enthielt jedes Szenario einen expliziten Hinweis darauf, dass Frau/ Herr K. sich angeblich „überraschend“ zu dem relevanten Ereignis äußern musste (keine Vorbereitungszeit) oder sich „gründlich überlegt“ was er/sie dazu sagen wird und „ausreichend Zeit hat seine/ ihre Antwort zu planen“ (mit Vorbereitungszeit).

Abschließend wurden Kontrollfragen vorgelegt, um den Erfolg der experimentellen Manipulation zu überprüfen. Vier Fragen bezogen sich auf die Wahrnehmung der Situation (z.B. „Wie verwerflich finden Sie es, dass Frau/ Herr K. in der geschilderten Situation lügt“) und eine auf die angenommene Erfolgswahrscheinlichkeit („Für wie wahrscheinlich halten Sie es, dass Frau/ Herr K. der Lüge überführt wird“). Alle Fragen waren auf 7-stufigen bipolaren Adjektivskalen von –3 bis +3 zu beurteilen. Die vermeintliche Vorbereitungsdauer wurde zum einen ebenfalls über eine 7-stufige bipolare Adjektivskala erfasst („Wie gut ist Frau/ Herr K. auf seine Lüge vorbereitet?“), zum anderen gaben die

Versuchspersonen eine freie Schätzung ab, wieviele Minuten Frau/Herr K. für die Vorbereitung ihrer/seiner Lüge aufgewendet hat.

Ergebnisse

Manipulation Checks

Die experimentelle Manipulation der Vorbereitung war erfolgreich. In den Bedingungen ohne Vorbereitung schätzten Versuchspersonen die Vorbereitungsdauer auf durchschnittlich 4.29 Minuten (SD = 5.23), in Bedingungen mit Vorbereitungszeit hingegen auf 38.63 Minuten (SD = 40.90), $F(1,228) = 101.96$, $p < .001$. Auch gingen sie davon aus, dass Herr/ Frau K. besser auf die Aussage vorbereitet war, wenn Gelegenheit zur Vorbereitung bestand ($M = 1.11$, SD = 1.62), als wenn er/sie gezwungen war, spontan auszusagen ($M = -1.25$, SD = 1.70), $F(1,228) = 140.99$, $p < .001$.

Die Analyse der situationsbezogenen Kontrollitems zeigte, dass die Verwerflichkeit der Lüge ($M = -0.05$, SD = 1,91), die Angst entdeckt zu werden ($M = 0.29$, SD = 1.57), die Motivation zu überzeugen ($M = 1.64$, SD = 1.17) und die negativen Konsequenzen bei einer Entdeckung der Lüge ($M = 0.41$, SD = 1.62) in der Verspätungssituation signifikant geringer eingeschätzt wurden als in den anderen beiden Situationen, alle Fs$(1,228) \geq 15.26$, alle ps $< .001$. Entgegen unserer Erwartungen differenzierten diese Kontrollitems jedoch nicht signifikant zwischen der Affären- ($M = 1.08$, SD = 1.64; $M = 2.07$, SD = 1.12; $M = 2.33$, SD = 0.95; bzw. $M = 2.21$, SD = 1.18) und Verbrechenssituation ($M = 1.24$, SD = 1.99; $M = 2.39$, SD = 1.16; $M = 2.45$, SD = 0.88; $M = 2.14$, SD = 1.47), alle Fs$(1,228) \leq 2{,}43$, alle ps $> .121$. Die Wahrscheinlichkeit, der Lüge überführt zu werden, wurde insgesamt hoch eingestuft, mit dem höchsten Wert in der Verbrechenssituation ($M = 1.98$, SD = 1.15), einem geringeren in der Affärensituation ($M = 0.79$, SD = 1.35) und dem geringsten in der Verspätungssituation ($M = - 0.23$, SD = 1.82), alle Fs$(1,228) \geq 20.12$, alle ps $< .001$.

Vergleich subjektiver und objektiver Täuschungsindikatoren

Anhand von one-sample *t*-Tests wurde geprüft, ob sich die Bewertungen der Versuchspersonen über alle sechs Experimentalbedingungen hinweg signifikant von Null unterscheiden. Weil dazu 73 statistische Tests durchzuführen waren, wurde ein Bonferroni-adjustiertes Entscheidungskriterium von $p < .0007$ festgelegt. Die Ergebnisse zeigten, von welchen der vorgegebenen Verhaltensweisen angenommen wurde, dass sie bei falschen Aussagen zu- bzw. abnehmen. In Tabelle 1 sind die resultierenden subjektiven Annahmen den von DePaulo et al. (2003) berichteten metaanalytischen Befunden zu objektiven Indikatoren gegenübergestellt. Um eine bessere Vergleichbarkeit zu ermöglichen, wurden die Mittelwerte der subjektiven Annahmen in das Effektstärkenmaß *d* transformiert. Insgesamt erwiesen sich 52 der 73 untersuchten Verhaltensweisen als subjektive Lügenindikatoren, obwohl nach den Befunden von DePaulo et al. (2003) nur 15 als objektive Lügenindikatoren zu werten sind. Zudem wurden sehr deutliche Verhaltensunterschiede erwartet. Bei den von uns untersuchten Indikatoren gingen die Versuchspersonen davon aus, dass sich deren Ausprägung bei Personen, die lügen und wahr aussagen, durchschnittlich um eine halbe Standardabweichung unterscheidet ($M_d = 0.54$, $SD_d = 0.39$, $Min_d = 0.00$, $Max_d = 1.42$). Die Gegenüberstellung zu den Befunden von DePaulo et al. (2003) zeigt jedoch, dass die Verhaltensunterschiede tatsächlich nur gering ausfallen ($M_d = 0.17$, $SD_d = 0.17$, $Min_d = 0.00$, $Max_d = 0.70$). Die klassischen Befunde zur Diskrepanz subjektiver und objektiver Täuschungsannahmen wurden demnach repliziert.

Moderatoren subjektiver Täuschungsindikatoren

Es wurde eine 3 x 2 MANOVA mit den Faktoren Situation (Verspätungs-, Affären-, Verbrechenssituation) und Vorbereitungszeit (ohne/mit Vorbereitung) gerechnet. Abhängige Variablen waren die Antworten der Versuchspersonen zu allen 73 Items.

Beide Faktoren erzielten einen signifikanten multivariaten Haupteffekt. Die subjektiven Annahmen zu Täuschungsindikatoren unterschieden sich sowohl in Abhängigkeit von der Vorbereitungszeit, Wilks' Lambda = .61, multivariates $F(73,162) = 1.44$, $p = .028$, als auch in Abhängigkeit von der Situation, Wilks' Lambda = .33, multivariates $F(146,324) = 1.65$, $p < .001$. Die Wechselwirkung der beiden Faktoren war nicht signifikant, Wilks' Lambda = .48, multivariates $F(146,324) = 0.97$, $p = .583$.

Zusatzanalysen mit dem Geschlecht der Stimulus- und der Versuchspersonen als zusätzliche Faktoren ergaben sowohl einen multivariaten Haupteffekt des Geschlechts der Versuchspersonen sowie eine Reihe signifikanter multivariater Wechselwirkungseffekte dieser beiden mit den anderen Faktoren. Daher sollten Geschlechterunterschiede bei zukünftigen Untersuchungen beachtet werden. Aufgrund der geringen Zellenbesetzung von $n = 10$ wird jedoch für die vorliegende Untersuchung auf eine Interpretation dieser Geschlechtereffekte verzichtet.

Im Folgenden werden die jeweiligen signifikanten univariaten Analysen für die Vorbereitungszeit und Täuschungssituation getrennt erläutert. Die univariaten Tests sind durch die signifikanten multivariaten Effekte abgesichert.

Vorbereitungszeit

Der Effekt der Vorbereitungszeit ließ sich univariat vor allem auf unterschiedliche subjektive Annahmen zu Indikatoren im Sprachbereich zurückführen. Die Versuchpersonen erwarteten, dass erlogene Aussagen mehr grammatikalische Fehler aufweisen und stockender vorgetragen werden als wahre. In Bedingungen mit Vorbereitungszeit wurden nur geringe, in Bedingungen ohne Vorbereitung Unterschiede mittlerer Größenordnung erwartet. Zudem wurde angenommen, dass Lügner weniger monoton sprechen, als Personen, die wahr aussagen. Dieser Effekt war in Bedingungen ohne Vorbereitungszeit signifikant stärker als in Bedingungen mit Vorbereitungszeit. In den Bedingungen ohne Vorbereitungszeit erwarteten die Versuchspersonen bei erlogenen Aussagen

einen geringeren Redeanteil und kürzere Sätze als bei wahren. Hingegen wurde in Bedingungen mit Vorbereitungszeit vermutet, dass Lügen mit einer deutlichen Zunahme des Redeanteils sowie tendenziell kürzeren Sätzen einhergehen. Zudem zeigte sich ein signifikanter univariater Haupteffekt der Vorbereitungszeit bezüglich der Ausführlichkeit der Antworten. In den Bedingungen ohne Vorbereitungszeit wurden keine Unterschiede in der Ausführlichkeit wahrer und erlogener Aussagen erwartet. Im Gegensatz dazu wurde angenommen, dass vorbereitete Lügner weniger ausführlich antworten als Personen, die wahr aussagen.

Von den Indikatoren im Gesichtsbereich wurde nur die Dauer des Blickkontaktes in Abhängigkeit von der Vorbereitungszeit unterschiedlich bewertet. Versuchspersonen erwarteten, dass Lügner generell kürzer Blickkontakt halten als Personen, die wahr aussagen. In Bedingungen mit Vorbereitung wurde angenommen, dass dieser Unterschied signifikant deutlicher ausgeprägt ist als in Bedingungen ohne Vorbereitungszeit. Bei den Indikatoren im Körperbereich waren die subjektiven Annahmen in den Bedingungen mit und ohne Vorbereitung vergleichbar. Bezogen auf den Gesamteindruck variierten die subjektiven Annahmen zu zwei Indikatoren in Abhängigkeit von der Untersuchungsbedingung. Es wurde erwartet, dass bei erlogenen Aussagen eine unangenehmere Gesprächsatmosphäre vorherrscht und die Person einen nervöseren Gesamteindruck hinterlässt als bei wahren Aussagen. Diese Unterschiede wurden in den Bedingungen ohne Vorbereitung signifikant stärker eingeschätzt als in Bedingungen mit Vorbereitung.

Täuschungssituation

Wenn sich ein univariater Effekt der Täuschungssituation zeigte, wurde anschließend geprüft, welche der drei Versuchsbedingungen sich signifikant voneinander unterschieden. Dazu wurden für jede abhängige Variable drei paarweise Vergleiche durchgeführt und gegen ein Bonferroni-adjustiertes Signifikanzniveau von $p < .017$ abgesichert.

Ein Effekt der Täuschungssituation zeigte sich bei zwei Indikatoren im Sprachbereich. Die Versuchspersonen beurteilten generell vermehrte Antwortverzögerungen und einen geringeren Redeanteil als Lügenindikatoren. Bezüglich der Antwortverzögerungen zeigte sich in der Affärensituation ein sehr starker Effekt, der sich signifikant von dem mittleren Effekt in der Verbrechenssituation unterschied. Bezüglich des Redeanteils unterschieden sich die beiden Alltagssituationen signifikant. In der Affärensituation wurde nur tendenziell ein geringerer Redeanteil bei lügenden im Vergleich zu wahr aussagenden Personen angenommen. Im Gegensatz dazu wurde in der Verspätungssituation ein Unterschied mittlerer Größenordnung erwartet.

Die subjektiven Annahmen variierten am häufigsten für die Täuschungsindikatoren im Gesichtsbereich. In Abhängigkeit von der vorgegebenen Situation erwarteten die Versuchspersonen unterschiedlich häufig Lachen, Grinsen und Lächeln bei Lügnern. Während die Versuchspersonen in der Verbrechenssituation erwarteten, dass diese Verhaltensweisen abnehmen, wurde in den beiden Alltagssittuationen angenommen, dass lügende Personen diese Verhaltensweisen häufiger zeigen als Personen, die wahr aussagen. Die entsprechenden Effektstärken waren jedoch nur gering (siehe Tabelle 1). Zudem wurde erwartet, dass lügende Personen häufiger nicken, als Personen die wahr aussagen. Dabei war der angenommene Verhaltensunterschied in der Verbrechenssituation signifikant stärker ausgeprägt als in der Affärensituation. In der Affärensituation wurde angenommen, dass Lügner häufiger den Kopf schütteln als Personen, die sich wahr äußern. Der Unterschied zur Verspätungssituation, in der keine Verhaltensunterschiede zwischen lügenden und wahr aussagenden Personen angenommen wurden, verfehlte jedoch knapp das festgelegte Signifikanzniveau. Schließlich unterschieden sich die subjektiven Annahmen in der Verspätungs- und Affärensituation bezogen auf den Indikator „blass werden im Gesicht“. Während in der Verspätungssituation erwartet wurde, dass Lügner seltener erblassen, wurde in der Verbrechenssituation erwartet,

dass sie häufiger erblassen als Personen, die wahr aussagen. Die Effektstärken waren jedoch nur gering.

Für die Indikatoren im Körperbereich zeigten sich wiederum keine Unterschiede in Abhängigkeit von der vorgegebenen Situation. Die subjektiven Annahmen unterschieden sich erneut in der Bewertung des Gesamteindrucks. In allen Bedingungen wurde erwartet, dass im Gespräch mit lügenden Personen eine deutlich unangenehmere Atmosphäre vorherrscht als im Gespräch mit wahr aussagenden. Dieser Effekt war für die Verbrechenssituation am stärksten ausgeprägt und unterschied sich signifikant von der Verspätungssituation. Zudem erwarteten die Versuchspersonen, dass Personen beim Lügen nachdenklicher wirken, als wenn sie wahrheitsgemäß aussagen. Diesbezüglich zeigte sich in der Verspätungssituation ein mittlerer und in der Affärensituation ein starker Effekt. Die Bewertung der Nachdenklichkeit in der Verbrechenssituation fiel zwischen die beiden Alltagssituationen, ohne sich signifikant von diesen zu unterscheiden.

Diskussion

Die vorliegende Studie untersuchte subjektive Annahmen über Täuschungsindikatoren in Abhängigkeit von der Gelegenheit zur Vorbereitung einer Aussage und dem Gegenstand der Lüge. Diese beiden Faktoren wurden über die Vorgabe fiktiver Szenarien variiert.

Die experimentelle Manipulation der Vorbereitungszeit war erfolgreich, während die Manipulation des Gegenstands der Lüge nur teilweise gelang. Wider Erwarten wurden die Affären- und Verbrechenssituation ähnlich wahrgenommen, weil die Affärensituation prekärer eingeschätzt wurde als erwartet. Möglicherweise lässt sich das darauf zurückführen, dass die beiden Situationen einen unterschiedlichen Bekanntheitsgrad zwischen den Interaktionspartnern implizierten. Zudem lässt sich spekulieren, dass die Versuchspersonen die von uns konzipierte Affärensituation extrem negativ beurteilten, weil sie ihnen vertrauter ist, während bei der Verbrechenssituation ein persönlicher Erfahrungsbezug unwahrscheinlich ist. Auch die Verwerflichkeit zu lügen wurde für beide Situationen

gleichermaßen hoch bewertet. Möglicherweise spiegelt die wenig extreme Beurteilung der Verbrechenssituation wider, dass zwar die Tötungshandlung an sich als verwerflich beurteilt wurde, jedoch weniger der Versuch, durch eine Lüge den strafrechtlichen Konsequenzen zu entgehen. Die Analyse der Kontrollitems zeigte aber auch, dass die Verspätungssituation erwartungsgemäß wahrgenommen wurde, so dass zumindest von einer zweifachen Abstufung der Situation ausgegangen werden kann. Insofern scheint es gerechtfertigt, die subjektiven Annahmen zu Täuschungsindikatoren auch in Abhängigkeit vom Gegenstand der Lüge zu betrachten.

Nur zehn der 73 untersuchten Verhaltensweisen wurden in Abhängigkeit vom Gegenstand der Lüge unterschiedlich bewertet. Beispielsweise wurde hinsichtlich des Gesamteindrucks erwartet, dass Personen nachdenklicher wirken und eine unangenehmere Gesprächsatmosphäre vorherrscht, wenn sie lügen als wenn sie wahr aussagen. Bezüglich der Nachdenklichkeit waren die subjektiven Annahmen für die Affärensituation signifikant deutlicher ausgeprägt als für die Verspätungssituation. Bezüglich der Gesprächsatmosphäre waren die Effekte für die Verbrechenssituation stärker. Scheinbar gehen Beurteiler davon aus, dass ernste Lügen mit mehr kognitiver Anstrengung und Stress einhergehen, jedoch keine anderen Verhaltensweisen hervorrufen als triviale Lügen (vgl. Vrij & Taylor, 2003).

In Abhängigkeit von der Gelegenheit zur Vorbereitung wurden neun Indikatoren unterschiedlich beurteilt. Vor allem im sprachlichen Bereich zeigten sich Unterschiede. So wurde in Bedingungen mit Vorbereitung erwartet, dass der Redeanteil von Lügnern deutlich geringer ist als von Personen, die wahr aussagen. Hingegen wurde in Bedingungen ohne Vorbereitung angenommen, dass lügende Personen im Vergleich zu wahr aussagenden mehr reden. Dies entspricht nicht den Befunden zu objektiven Indikatoren (Sporer & Schwandt, 2006). Weiterhin wurde angenommen, dass erlogene Aussagen mehr grammatikalische Fehler aufweisen, stockender und monotoner vorgetragen

werden als wahre, vor allem in den Bedingungen ohne Vorbereitung. Auch die subjektive Annahme eines nervöseren Gesamteindrucks lügender im Vergleich zu wahr aussagenden Personen war in Bedingungen ohne Vorbereitung deutlicher ausgeprägt als in Bedingungen mit Vorbereitung. Es zeigten sich jedoch keine unterschiedlichen Bewertungen bei den Verhaltensweisen im Körperbereich. Das ist überraschend, weil diese Kategorie vor allem Bewegungsverhalten umfasst, das allgemein mit Nervosität assoziiert wird. Obwohl Personen annahmen, dass unvorbereitete Lügner noch nervöser wirken als vorbereitete, spiegelte sich dies also nicht in einer durchgängig entsprechenden Bewertung einzelner Verhaltensweisen wider.

Insgesamt variierten die subjektiven Annahmen in Abhängigkeit von der Vorbereitungszeit und der Situation für einzelne Täuschungsindikatoren. Dies betrifft im Gegensatz zu den von Taylor und Vrij (2000) berichteten Ergebnissen auch non- und paraverbale Verhaltensweisen. Teilweise lassen sich die divergierenden Befunde darauf zurückführen, dass in unserer Untersuchung spezifischere Verhaltensweisen abgefragt wurden. Beispielsweise erfaßten Taylor und Vrij subjektive Annahmen zu Kopfbewegungen und Lächeln, während in der vorliegenden Studie zwischen „Kopfnicken“ und „Kopfschütteln“ beziehungsweise zwischen „Lachen“, „Grinsen“ und “Lächeln“ differenziert wurde. Zudem konnten Taylor und Vrij (2000) keinen signifikanten Effekt der kognitiven Komplexität nachweisen. Die Autoren räumten selbst ein, dass durch ihre Art der Operationalisierung möglicherweise auch die Motivation erfolgreich zu lügen beeinflusst wurde. In der vorliegenden Studie wurde die Aussagekomplexität über die Gelegenheit zur Vorbereitung variiert. Diese Variation wirkte sich beispielsweise auf die subjektiven Annahmen bezüglich der Ausführlichkeit von Antworten, des Blickkontaktes und der Redeweise aus.

Schließlich zeigten sich auch in unserer Untersuchung typische Diskrepanzen zwischen subjektiven und objektiven Täuschungsindikatoren, beispielsweise in der Annahme, Lügner würden ihren Blick abwenden, obwohl

dies tatsächlich kein Lügenindikator ist (vgl. DePaulo et al., 2003). Ebenso wurden erneut stärkere Unterschiede im Verhalten von lügenden und wahr aussagenden Personen angenommen, als objektiv nachweisbar sind.

Praktische Implikationen und weiterführender Forschungsbedarf

Die Vorbereitungszeit und der Gegenstand der Lüge haben sich als relevante Moderatoren für objektive Täuschungsindikatoren erwiesen (z.B. Greene, O´Hair, Cody, & Yen, 1985; Miller, DeTurck, & Kalbfleisch, 1983; Sporer & Schwandt, 2006, 2007). Unsere Untersuchungsbefunde weisen darauf hin, dass ihnen in Bezug auf subjektive Täuschungsannahmen eine vergleichsweise geringere Bedeutung zukommt. Laien scheinen bei der Beurteilung von Lügenindikatoren den Kontext einer Aussage nicht in dem Maße zu berücksichtigen, wie er tatsächlich bedeutsam ist. Beispielsweise berichteten DePaulo et al. (2003), dass das Ausmaß an Blickkontakt, Fuß- und Beinbewegungen, die Nervosität im Gesamteindruck und die Stimmlage in Studien, bei denen motivationale Anreize zu lügen geschaffen wurden, objektiv zwischen wahren und erlogenen Aussagen differieren. Zudem zeigte sich eine verlängerte Antwortlatenz von Lügnern im Vergleich zu Personen, die wahr aussagen, wenn keine Gelegenheit bestand, die Aussage vorzubereiten. In den subjektiven Annahmen unserer Versuchspersonen gab es keine vergleichbaren Unterschiede.

Die praktischen Implikationen unserer Befunde beziehen sich auf Möglichkeiten zur Förderung der Beurteilungsfähigkeit. Im Rahmen von Trainings könnten Beurteiler zu einer Analyse der genauen Begleitumstände einer Aussage angeleitet werden. Sie sollten sich fragen, ob eine Person Gelegenheit hatte, ihre Aussage vorzubereiten, und wie viel sie mit einer Lüge riskiert. Ebenso könnte es sinnvoll sein, Beurteiler für weitere Moderatoren objektiver Täuschungsindikatoren zu sensibilisieren, beispielsweise die Komplexität einer Aussage, die Nachprüfbarkeit von Informationen usw.. Auch auf die Möglichkeit intra-individueller

Vergleiche (z.B. zwischen Fragen zu eher neutralen versus „tatrelevanten“ Themen) sollte hingewiesen werden.

Jedoch sind weitere Forschungsbemühungen unerlässlich, um die vorliegenden Befunde abzusichern und zu ergänzen. So sollten bei der Untersuchung subjektiver Täuschungsindikatoren weitere Aspekte variiert werden, die sich als relevante Moderatoren objektiver Täuschungsindikatoren erwiesen haben. Dabei sollten auch Wechselwirkungen mit Geschlechtereffekten berücksichtigt werden. Nach neueren Forschungsbefunden weisen inhaltliche Aussagemerkmale eine bessere objektive Differenzierungskraft auf als nonverbale und paraverbale Indikatoren (DePaulo et al., 2003). Daher sollte auch für inhaltliche Aussagemerkmale überprüft werden, ob die subjektiven Täuschungsannahmen in Abhängigkeit von Moderatorvariablen variieren. Zudem wäre es interessant die subjektiven Täuschungsannahmen bei vertrauten Personen entsprechend zu untersuchen.[1] So ließe sich spekulieren, dass Personen, wenn sie das Verhalten Ihrer Partner in verschiedenen Kontexten kennen, auch annehmen, dass deren Lügenverhalten in Abhängigkeit von kontextuellen Faktoren variieren kann.

Insgesamt erscheint es in Anbetracht der berufsübergreifend unzureichenden Fähigkeit, falsche von wahren Aussagen zu unterscheiden, wichtig, ein differenzierteres Verständnis subjektiver Täuschungsannahmen zu erlangen. Dies könnte wesentlich dazu beitragen, effektivere Trainings zur Förderung der Entdeckung von Täuschung zu entwickeln.

---

[1] Wir danken einem anonymen Reviewer für diese Anregung.

Literatur

Akehurst, L., Köhnken, G., Vrij, A., & Bull, R. (1996). Lay persons' and police officers' beliefs regarding deceptive behaviour. Applied Cognitive Psychology, 10, 461-471.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale: Lawrence Erlbaum.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. Psychological Bulletin, 129, 74-112.

DePaulo, B. M., Stone, J. L., & Lassiter, G. D. (1985). Deceiving and detecting deceit. In B. R. Schenkler (Ed.), The self and social life (pp. 323-370). New York: McGraw-Hill.

Ekman, P. (1992). Telling lies: Clues to deceit in the marketplace, politics and marriage. New York: W. W. Norton.

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? American Psychologist, 46, 913-920.

Fiedler, K. (1989). Lügendetektion aus alltagspsychologischer Sicht. Psychologische Rundschau, 40, 127-140.

Greene, J. O., O´Hair, H. D., Cody, M. J., & Yen, C. (1985). Planning and control of behavior during deception. Human Communication Research, 11, 335-364.

Hocking, J. E., & Leathers, D. G. (1980). Nonverbal indicators of deception: A new theoretical perspective. Communication Monographs, 47, 119-131.

Köhnken, G. (1988). Glaubwürdigkeit: Empirische und theoretische Untersuchungen zu einem psychologischen Konstrukt. Kiel: Unveröffentlichte Habilitationsschrift.

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

Miller, G. R., DeTurck, M. A., & Kalbfleisch, P. J. (1983). Self-Monitoring, rehearsal, and deceptive communication. Human Communication Research, 10, 97-117.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231-259.

Reinhard, M. A., Burghardt, K., Sporer, S. L., & Bursch, S. E. (2002). Alltagsvorstellungen über inhaltliche Kennzeichen von Lügen: Selbstberichtete Begründungen bei konkreten Glaubürdigkeitsurteilen. Zeitschrift für Sozialpsychologie, 33, 169-180.

Sporer, S. L. (1997). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. Applied Cognitive Psychology, 11, 373-397.

Sporer, S. L. (2004, June). Evaluating eyewitness testimony. Paper presented at the Max-Planck-Institute Conference on Judicial Representation in Bad Seeon, Germany.

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. Psychology, Public Policy, and Law, 13, 1-34.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. Applied Cognitive Psychology, 20, 421-446.

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

Strömwall, L. A., & Granhag, P. A. (2003). How to detect deception? Arresting the beliefs of police officers, prosecutors and judges. Psychology, Crime, and Law, 9, 19-36.

Taylor, R., & Vrij, A. (2000). The effects of varying stake and cognitive complexity on beliefs about the cues to deceeption. International Journal of Police Science and Management, 3, 111-123.

Volbert, R., & Busse, D. (1995). Wie fair sind Verfahren für kindliche Zeugen? Zur Strafverfolgung bei sexuellem Mißbrauch von Kindern. In G. Bierbrauer, W. Gottwald, & B. Birnbreier-Stahlberger (Eds.), Verfahrensgerechtigkeit: Rechtspsychologische Forschungsbeiträge für die Justizpraxis (pp. 139-162). Köln: Verlag Dr. Otto Schmidt KG.

Vrij, A. (2000). Detecting lies and deceit: The psychology of lying and implications for professional practice. Chichester: John Wiley.

Vrij, A., & Taylor, R. (2003). Police officers' and students' beliefs about telling and detecting trivial and serious lies. International Journal of Police Science and Management, 5, 41-49.

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.

Zuckerman, M., Koestner, R., & Driver, R. (1981). Beliefs about cues associated with deception. Journal of Nonverbal Behavior, 6, 105-114.

STUDIE 2

Subjektive Indikatoren von Täuschung: Alltagsvorstellungen zu inhaltlichen Aussagemerkmalen

Die Unterscheidung wahrer und erfundener Aussagen ist in privaten, beruflichen und juristischen Kontexten von Bedeutung. Allerdings wurde wiederholt gezeigt, dass es Personen kaum gelingt den Wahrheitsstatus von Aussagen richtig einzuschätzen. Nach einer aktuellen Metaanalyse von Bond und DePaulo (2006) liegt die mittlere Urteilsrichtigkeit bei 54%, also nur geringfügig über dem Zufallsniveau (bei einer Basisrate von 50%, d.h. wenn gleich viele tatsächlich erfundene und wahre Aussagen beurteilt werden). Dies gilt nicht nur für Laien, sondern auch für Personen, die berufsbedingt zwischen wahren und erfundenen Aussagen differenzieren sollen, wie zum Beispiel Polizisten, Detektive, Psychologen und Richter (Aamodt & Custer, 2006).

Die Forschung hat sich einerseits damit beschäftigt, tatsächliche Unterschiede zwischen wahren und erfundenen Aussagen nachzuweisen (objektive Indikatoren). Andererseits wurde untersucht, von welchen Verhaltensweisen Personen annehmen, dass sie unterschiedlich häufig bei wahren und erfundenen Aussagen vorzufinden seien (subjektive Indikatoren). Die Identifikation subjektiver Täuschungsindikatoren ist aus zwei Gründen von Interesse. Zum einen sollten für eine erfolgreiche Lügendetektion objektive und subjektive Indikatoren weitestgehend übereinstimmen. Diskrepanzen verweisen darauf, dass Personen falsche Annahmen darüber haben, welche Verhaltensweisen mit wahren und erfundenen Aussagen assoziiert sind. Dies liefert eine mögliche Erklärung für die schlechte Fähigkeit Lügen zu entdecken. Zum anderen wird im Rahmen des sogenannten Kontrollansatzes davon ausgegangen, dass Lügner gezielt Verhaltensweisen vermeiden, von denen sie annehmen, dass sie mit Täuschung assoziiert seien (Zuckerman, DePaulo & Rosenthal, 1981). Entsprechend wurden Glaubhaftigkeitsmerkmale entwickelt,

deren Validität darüber begründet wird, dass sie stereotypen Vorstellungen von wahren Aussagen widersprechen und deswegen bei Falschaussagen gezielt vermieden werden. Daher gilt es nachzuweisen, dass diese Merkmale subjektiv als Täuschungsindikatoren aufgefasst werden.

Gegenstand der vorliegenden Untersuchung sind subjektive Annahmen zu inhaltlichen Aussagemerkmalen. Inhaltliche Merkmale sind abzugrenzen von nonverbalen Merkmalen, wie Mimik und Gestik, die vollkommen unabhängig von der Aussage selbst beobachtbar sind, sowie von paraverbalen Merkmalen, die sich auf das Sprechverhalten beziehen (zu subjektiven Annahmen bezüglich nonverbaler und paraverbaler Merkmale s. Breuer, Sporer & Reinhard, 2005; Köhnken, 1990; Strömwall, Granhag & Hartwig, 2004; Vrij, 2000/2008).

Inhaltliche Aussagemerkmale

Eine Vielzahl von inhaltlichen Aussagemerkmalen wurde auf ihre objektive Differenzierungskraft wahrer und erfundener Aussagen hin untersucht, d.h. es wurde überprüft, ob sie tatsächlich unterschiedlich häufig bei wahren und erfundenen Aussagen vorzufinden sind. Insbesondere die Merkmale der sogenannten Criteria-Based Content Analysis (CBCA) haben dabei breites Forschungsinteresse erfahren. Diese gehen auf eine Systematisierung von Steller und Köhnken (1989) zurück. Dabei sind fünf Kategorien zu unterscheiden, denen insgesamt 19 Glaubhaftigkeitsmerkmale zugeordnet sind. Es wird postuliert, dass sämtliche CBCA-Merkmale eher bei wahren als bei erfundenen Aussagen vorzufinden sind. Die „allgemeinen Merkmale“ beziehen sich auf die Aussage als Ganzes und werden als notwendige, aber keinesfalls hinreichende Bedingungen erlebnisbasierter Aussagen aufgefasst (Steller, 1989, S. 136). Hingegen ist es zur Beurteilung der „speziellen Inhalte“ und „inhaltlichen Besonderheiten“ erforderlich, einzelne Teile der Aussage genauer zu betrachten. Die Validität dieser beiden Merkmalskategorien führen Steller und Köhnken auf kognitive Faktoren zurück. Wahre Aussagen sollten eine höhere inhaltliche Qualität als falsche aufweisen,

weil angenommen wird, dass einzelne Merkmale schwierig zu erfinden sind. Die Kategorie der sogenannten „motivationsbezogenen Inhalte“ umfasst Merkmale, von denen angenommen wird, dass sie stereotypen Vorstellungen von Wahrheit widersprechen. Entsprechend bezeichneten Ruby und Brigham sie auch als „contrary-to-stereotype contents“ (1998, S. 370). Unter der Voraussetzung, dass falsch aussagende Personen in besonderem Maße um eine positive Selbstdarstellung bemüht sind (DePaulo, Lindsay, Malone, Muhlenbruck, Charlton & Cooper, 2003; Fiedler, 1989a), sollten sie diese Merkmale gezielt vermeiden. Die von Steller und Köhnken (1989) gewählte Bezeichnung “motivation-related contents” wurde häufig übernommen bzw. wörtlich übersetzt (z.B. Steller, Wellershaus & Wolf, 1992). Der Verweis auf motivationale Prozesse ist jedoch differenziert zu betrachten. Es wird lediglich vorausgesetzt, dass Personen darum bemüht sind, glaubhaft und aufrichtig zu erscheinen (vgl. Fiedler, 1989b). Handlungsleitend sind jedoch stereotype Vorstellungen über Täuschungsindikatoren und damit spezifische Kognitionen. Daher werden die entsprechenden Merkmale im Folgenden in Anlehnung an Ruby und Brigham (1998) als stereotypkonträr bezeichnet. Innerhalb der letzten Kategorie werden „deliktspezifische Inhalte“ berücksichtigt, deren Vorhandensein ebenfalls die Hypothese eines persönlichen Erfahrungsbezugs stärken kann.

Wiederholt wurde die unzureichende theoretische Untermauerung der CBCA-Merkmale kritisiert (z.B. Steller & Köhnken, 1989; Sporer, 1997a). Daher gab es Bemühungen, aus dem sogenannten Realitätsüberwachungsansatz (RÜ, Johnson & Raye, 1981) Aussagemerkmale für die Glaubhaftigkeitsanalyse abzuleiten und deren Validität zu überprüfen. Vor dem Hintergrund des RÜ-Ansatzes sind bei wahren Aussagen mehr Hinweise auf räumliche, zeitliche und sensorische Details als bei erfundenen zu erwarten. Zudem wird postuliert, dass erfundene Aussagen mehr Hinweise auf kognitive Operationen enthalten als wahre.

Sporer (1996/1998/2004) wiederum hat mit den Aberdeen Report Judgment Scales (ARJS) theoretisch abgeleitete Aussagemerkmale vorgestellt, die sowohl CBCA- und RÜ-Merkmale integrieren, als auch sozialpsychologische und gedächtnistheoretische Aspekte, insbesondere zu autobiographischen Erinnerungen, berücksichtigen. Erste Ergebnisse zur Reliabilität und Validität der ARJS liegen bereits vor (vgl. Sporer, 2004, für eine Übersicht). Diese dienten als Grundlage der vorliegenden Untersuchung.

Objektive Differenzierungskraft inhaltlicher Aussagemerkmale

DePaulo et al. (2003) fassten den umfangreichen Forschungsstand zu objektiven Indikatoren von Täuschung metaanalytisch zusammen. Den Befunden zufolge erlauben es inhaltliche Aussagemerkmale besser als nonverbale Indikatoren zwischen wahren und erfundenen Aussagen zu differenzieren.

DePaulo et al.'s (2003) Metaanalyse umfasste 16 CBCA-Merkmale. Allerdings war die Datenbasis zum Zeitpunkt ihrer Analysen noch gering, so dass die berichteten Effektstärken auf den Befunden von $\underline{k} \leq 6$ unabhängigen Stichproben basierten. Die CBCA-Merkmale wurden als Glaubhaftigkeitsmerkmale konzipiert, d.h. sie sollten häufiger bei wahren als bei erfundenen Aussagen vorzufinden sein. Nach den Befunden von DePaulo et al. fanden sich jedoch nur die Merkmale logische Konsistenz ($\underline{d} = 0.23$)[1], spontane

[1] DePaulo et al. (2003) berichteten negative Effektstärkemaße $\underline{d}$, wenn ein Merkmal seltener bei erfundenen als bei wahren Aussagen vorzufinden war. Um die Vergleichbarkeit mit der vorliegenden Untersuchung zu erleichtern, wurden die positiven und negativen Vorzeichen der zitierten Effektstärken ausgetauscht. Infolgedessen verweisen positive Effektstärken auf einen Wahrheitsindikator, d.h. das Merkmal wurde häufiger bei wahren als bei erfundenen Aussagen vorgefunden. Negative Effektstärken indizieren hingegen, dass ein Lügenindikator

Verbesserungen der eigenen Aussage (d = 0.29) sowie Eingeständnisse von Erinnerungslücken (d = 0.42) signifikant häufiger bei wahren als bei erfundenen Aussagen. Zudem berichteten die Autoren, dass wahre Aussagen detaillierter (d = 0.30) waren als erfundene, wobei dieses Aussagemerkmal globaler operationalisiert wurde als das CBCA-Kriterium quantitativer Detailreichtum. Raum-zeitliche Verknüpfungen (d = 0.21), Schilderungen ausgefallener Einzelheiten (d = 0.16) sowie Einwände gegen die Richtigkeit der eigenen Aussage (d = 0.10) waren jedoch zumindest tendenziell bei wahren Aussagen höher ausgeprägt als bei erfundenen. Wider Erwarten waren bei erfundenen Aussagen signifikant häufiger indirekt handlungsbezogene Schilderungen (d = -0.35) sowie tendenziell häufiger Selbstbelastungen (d = -0.21) und Schilderungen psychischer Vorgänge des Täters (d = -0.22) als bei wahren Aussagen vorzufinden. Für die Merkmale unstrukturierte Darstellung (d = -0.06), Schilderungen von Interaktionen (d = 0.03), Komplikationen im Handlungsverlauf (d = 0.04), nebensächlichen Einzelheiten (d = 0.01) sowie eigene psychische Vorgänge (d = -0.02) wurden keine Unterschiede zwischen erfundenen und wahren Aussagen festgestellt. Aus den Befunden für die beiden CBCA-Merkmale phänomengemäße Schilderung unverstandener Handlungselemente (d = 0.22) und Entlastung des Angeschuldigten (d = 0.00) lassen sich keine generellen Schlussfolgerungen ziehen, da ihnen lediglich die Ergebnisse von zwei bzw. einer Untersuchung zugrunde lagen.

DePaulo et al. (2003) beschränkten sich auf Primärstudien, in denen Aussagen von Erwachsenen untersucht wurden, und beendeten ihre Literaturrecherche im Oktober 1999. Vrij (2005) lieferte einen aktuelleren Forschungsüberblick zur Validität der CBCA, bei dem auch Untersuchungen kindlicher Aussagen berücksichtigt wurden. Die Ergebnisse von fünf Feld- und 17

---

vorliegt, d.h. das Merkmal wurde häufiger bei erfundenen als bei wahren Aussagen vorgefunden.

Laborstudien qualitativ zusammenfassend, schlussfolgerte er, dass die stereotypkonträren CBCA-Merkmale weniger Unterstützung erfahren haben als die allgemeinen Merkmale, speziellen Inhalte und inhaltlichen Besonderheiten. Auch Vrij berichtete, dass wahre Aussagen mehr Details umfassten, unstrukturierter dargestellt wurden und mehr Raum-zeitliche Verknüpfungen beinhalteten als erfundene. Zudem schlussfolgerte er aus seiner Literaturübersicht auf die erwartungsgemäße Differenzierungskraft der Wiedergabe von Gesprächen. Allerdings fand Vrij im Gegensatz zu den Befunden von DePaulo et al. in den meisten Studien keine signifikanten Unterschiede zwischen wahren und erfundenen Aussagen hinsichtlich des Eingeständnisses von Erinnerungslücken. Unterschiede zwischen wahren und erfundenen Aussagen im Ausmaß an Selbstbelastungen fielen in den von Vrij aufgeführten Studien erwartungskonträr aus. Der Befund, dass erfundene Aussagen mehr Selbstbelastungen aufwiesen, deckt sich jedoch mit den metaanalytischen Befunden von DePaulo et al.

DePaulo et al. (2003) integrierten auch sechs RÜ-Merkmale in ihre Metaanalyse. Ihre Analysen verwiesen darauf, dass wahre Aussagen mehr sensorische Details aufwiesen als erfundene ($\underline{d}$ = 0.17). Allerdings lagen nicht genügend Primärstudien vor, um die Validität der übrigen RÜ-Kriterien zuverlässig abschätzen zu können. Daher sind neuere qualitative Forschungsüberblicke zur Validität der RÜ-Merkmale aufschlussreicher (Masip, Sporer, Garrido & Herrero, 2005; Sporer, 2004). Nach der Literaturdurchsicht von Masip et al. (2005) liegen für sensorische Details widersprüchliche Forschungsbefunde vor. Hingegen interpretierten die Autoren die Forschungsbefunde für die RÜ-Kriterien kontextuelle, räumliche und zeitliche Informationen sowie Realismus als vielversprechend. Diese Merkmale wurden den Vorhersagen des RÜ-Ansatzes entsprechend oftmals häufiger bei wahren als bei erfundenen Aussagen vorgefunden. Im Gegensatz dazu sollten erfundene Aussagen dem RÜ-Ansatz zufolge mehr Hinweise auf kognitive Operationen beinhalten als wahre. Die Forschungsbefunde hierzu sind allerdings ernüchternd. In den meisten Studien

differenzierten kognitive Operationen nicht signifikant zwischen wahren und erfundenen Aussagen (Alonso-Quecuty, 1992; Granhag, Strömwall, Landström, 2006; Hernandez-Fernaud & Alonso-Quecuty, 1997; Sporer, 1997a, Sporer & Küpper, 1995; Sporer & Sharman, 2006; Strömwall et al., 2004). In einigen Untersuchungen erwiesen sie sich sogar entgegen der ursprünglichen Annahmen als Wahrheitskriterium (Sporer & Küpper, 2004; Vrij, Edward, Roberts & Bull, 2000). Dies gilt auch für Untersuchungen, die gezielt die Validität der ARJS überprüften (vgl. die unter Tabelle 2.3 aufgeführten Effektstärken der objektiven Indikatoren). Hingegen waren kognitive Operationen nur in zwei anderen Studien hypothesenkonform als Lügenkriterium zu werten (Vrij, Akehurst, Soukara & Bull, 2004a, 2004b). Die heterogenen Forschungsbefunde sind möglicherweise auf Unterschiede in der Operationalisierung dieses Aussagemerkmals, in den Untersuchungsdesigns und dem Stimulusmaterial, oder auch auf die Wirksamkeit von Moderatoren zurückzuführen.

Auch zu den ARJS liegen mittlerweile mehrere Validierungsstudien vor (Barnier, Sharman, McKay & Sporer, 2005; Sporer, 1998; Sporer & Burghardt, 2004; Sporer, Samweber & Stucke, 2000; Sporer & Walther, 2006). Die ARJS-Merkmale wurden erwartungsgemäß eher in wahren als in erfundenen Aussagen vorgefunden. In allen Studien erhielten wahre Aussagen signifikant oder tendenziell höhere Werte als erfundene hinsichtlich der Skalen Details, Memorieren und Gedächtnis sowie Emotionen und Gefühle. Zudem beinhalteten wahre Aussagen meist mehr außergewöhnliche Details als erfundene. Die Skalen Realismus und logische Struktur sowie räumliche Details und Sinneseindrücke differenzierten hingegen oftmals nicht signifikant zwischen wahren und erfundenen Aussagen. Dies lässt sich vermutlich darauf zurückführen, dass diese Merkmale sehr häufig oder sehr selten im Untersuchungsmaterial vorzufinden waren, also Decken- und Bodeneffekte vorlagen (Barnier et al., 2005; Sporer, 1998; Sporer & Walther, 2006).

Zusammenfassend haben sich einige Glaubhaftigkeitsmerkmale als valide erwiesen, wobei andere selten untersucht wurden und/oder keinerlei Differenzierungskraft aufwiesen. Insbesondere hinsichtlich der Validität von stereotypkonträren Aussagemerkmalen ergibt sich weiterer Klärungsbedarf. Daher erscheint es notwendig, die postulierte Korrespondenz zwischen diesen Inhaltsmerkmalen und den subjektiven Annahmen zu Täuschungsindikatoren zu überprüfen.

Moderatoren der objektiven Differenzierungskraft

Metaanalytisch wurde nachgewiesen, dass die objektive Differenzierungskraft einzelner non- und paraverbaler Indikatoren dem Einfluss von Moderatoren unterliegt (DePaulo et al., 2003; Sporer & Schwandt, 2006, 2007). Für inhaltliche Aussagemerkmale wurde die Wirksamkeit von Moderatoren bisher nicht metaanalytisch untersucht. Jedoch weisen die Befunde einzelner Studien darauf hin, dass das Ausmaß an Vorbereitungszeit die Differenzierungskraft inhaltlicher Aussagemerkmale beeinflusst (Alonso-Quecuty, 1992; Sporer, 1997a, 1998; Sporer & Küpper, 1995).

In einer Untersuchung von Suengas und Johnson (1988) wirkte sich bereits die Instruktion, selbst-erlebte und vorgestellte Ereignisse für jeweils 15 Sekunden gedanklich zu wiederholen (memorieren), auf die Selbstbeurteilung von Erinnerungsqualitäten anhand von RÜ-Merkmalen aus. Das Ausmaß an Klarheit, sensorischen Informationen sowie Gedanken und Gefühlen wurde für nicht-memorierte Ereignisse geringer beurteilt als für memorierte. Auch Sporer und Küpper (2004) wiesen für die Selbstbeurteilung von komplexen Aussagen einen Effekt der Vorbereitung nach. Es ergaben sich signifikante Unterschiede zwischen vorbereiteten und unvorbereiteten Aussagen für die RÜ-Merkmale Klarheit und Lebendigkeit sowie für zeitliche Informationen. Nach einer Woche Vorbereitungszeit waren diese Merkmale geringer ausgeprägt als bei spontanen Berichten. Zudem erwarteten Sporer und Küpper eine Wechselwirkung zwischen

dem Wahrheitsstatus von Aussagen und der Gelegenheit zur Vorbereitung auf die inhaltlichen Aussagemerkmale. Sie argumentierten, dass Unterschiede zwischen wahren und falschen Aussagen reduziert wären, wenn Personen die Gelegenheit nutzen, ihre erfundenen Aussagen inhaltlich anzureichern. Es sei jedoch ebenso denkbar, dass dies einen persönlichen Erfahrungsbezug voraussetzt. Dementsprechend würden nur Personen, die über Selbst-Erlebtes berichten von der Gelegenheit zur Vorbereitung profitieren und die qualitativen Unterschiede zwischen wahren und erfundenen Aussagen wären verstärkt. Sporer und Küpper konnten jedoch keine signifikante Wechselwirkung zwischen dem Wahrheitsstatus und der Vorbereitungszeit feststellen. Die Autoren diskutierten, dass die einwöchige Gelegenheit zur Vorbereitung von den Probanden möglicherweise nicht genutzt wurde.

Sporer und Burghardt (2004) variierten die Vorbereitungszeit, indem sie den Probanden entweder 2 oder 15 Minuten zur Planung ihrer Aussage einräumten. Die transkribierten Aussagen wurden später von zwei trainierten Ratern hinsichtlich der ARJS beurteilt. Bei dieser Art der Operationalisierung wurden bei 15-minütiger Vorbereitung mehr inhaltliche Aussagemerkmale vorgefunden als bei kurzer Vorbereitungszeit. Es zeigte sich jedoch erneut keine Wechselwirkung zwischen dem Wahrheitsstatus und der Vorbereitungszeit auf die Aussagequalität.

In anderen Untersuchungen wurden jedoch Wechselwirkungen zwischen dem Wahrheitsstatus von Aussagen und der Gelegenheit zur Vorbereitung festgestellt (Alonso-Quecuty, 1992; Sporer & Küpper, 1995; Sporer, 1997a; Sporer, 1998). Allerdings bleibt unklar, ob die Differenzierungskraft der inhaltlichen Aussagemerkmale durch Vorbereitung erhöht oder reduziert wird. So waren die qualitativen Unterschiede zwischen erfundenen und wahren Aussagen in den Untersuchungen von Sporer und Küpper (1995) sowie von Sporer (1997a) in Bedingungen ohne Vorbereitung weniger deutlich als in Bedingungen mit Vorbereitung. Im Gegensatz dazu berichteten Alonso-Quecuty (1992) sowie Sporer (1998) eine deutlichere Differenzierungskraft inhaltlicher Aussagemerkmale in

Bedingungen ohne Vorbereitung. Nicht zuletzt gibt es Hinweise darauf, dass unterschiedliche Arten der Vorbereitung differentielle Effekte auf die inhaltliche Aussagequalität haben können (z.B. Johnson & Suengas, 1989; Suengas & Johnson, 1988).

Für Anwendungszwecke erscheint es sinnvoll, bei der Untersuchung objektiver Täuschungskorrelate nicht nur die Vorbereitung, sondern auch die Motivation erfolgreich zu überzeugen zu berücksichtigen (Sporer & Burghardt, 2004; Sporer & Schwandt, 2006, 2007). Es ist davon auszugehen, dass Personen in juristischen Kontexten wesentlich stärker motiviert sind zu überzeugen als in Alltagssituationen. Allerdings ist es im Labor kaum möglich Bedingungen herzustellen, die eine ähnlich starke Motivation induzieren, wie sie bei Aussagen mit potenziell rechtlichen Konsequenzen zu erwarten ist. Eine Feldstudie, die diesbezüglich eine vergleichsweise hohe externe Validität aufweist, wurde von Vrij und Mann (2001) vorgestellt. Dabei wurde das Verhalten eines Mörders bei wahren und erlogenen Aussageelementen analysiert. Die Autoren beschränkten sich jedoch auf nonverbale Verhaltensunterschiede. Mann, Vrij und Bull (2002) analysierten Videoaufzeichnungen der polizeilichen Vernehmungen von 16 Tatverdächtigen, die verschiedener Delikte beschuldigt wurden. Doch auch dieses Stimulusmaterial wurde lediglich verwendet, um non- und paraverbale Indikatoren zu untersuchen (Mann et al., 2002) oder die Fähigkeit von Polizeibeamten wahre und erfundene Aussagen zu unterscheiden (Mann & Vrij, 2006; Mann, Vrij & Bull, 2004; Vrij, Mann, Robbins & Robinson, 2006). Entsprechende Untersuchungen zu inhaltlichen Aussagemerkmalen sind uns nicht bekannt. Eine Studie von Feeley und deTurck (1998) verweist jedoch darauf, dass bei der in Laborstudien häufig verwendeten Instruktion zu lügen (sanktionierte Lügen) andere Aussagemerkmale festzustellen sind, als wenn sich Personen eigenständig dazu entscheiden falsch auszusagen (unsanktionierte Lügen). Insgesamt wurde die potenzielle Moderatorwirkung der Relevanz zu lügen für inhaltliche Aussagemerkmale bislang nur unzureichend untersucht.

Forschungsparadigmen zur Erfassung subjektiver Annahmen

Zur Erfassung subjektiver Annahmen lassen sich mehrere Forschungsansätze unterscheiden. Im Rahmen von Fragebogenuntersuchungen geben Versuchspersonen für vorgegebene Indikatoren an, ob sie diese eher bzw. stärker ausgeprägt in wahren oder erfundenen Aussagen erwarten (Akehurst, Köhnken, Vrij & Bull, 1996; Forrest, Feldman & Tyler, 2004; Granhag, Andersson, Strömwall & Hartwig, 2004; Granhag, Strömwall & Hartwig, 2005; Strömwall & Granhag, 2003; Vrij, Akehurst & Knight, 2006; Vrij, Edward & Bull, 2001; Zuckerman, Koestner & Driver, 1981). Allerdings ist bei solchen Befragungen nicht auszuschließen, dass die Versuchpersonen ihre Angaben auf unterschiedliche Täuschungssituationen beziehen. Dies gilt insbesondere für Vergleiche zwischen Berufsgruppen und lässt sich durch eine Instruktion entweder an ernsthafte oder an triviale Lügen zu denken, ohne diese jedoch genauer zu spezifizieren (vgl. Vrij & Taylor, 2003), vermutlich nicht beheben. Daher geben neuere Untersuchungen fiktive Szenarien vor, auf die sich die subjektiven Annahmen zu Täuschungskorrelaten beziehen sollen (Breuer et al., 2005; Lakhani & Taylor, 2003; Taylor & Hill-Davies, 2004; Taylor & Vrij, 2000). Taylor und Hick (2007) modifizierten dieses Vorgehen, indem sie die Probanden entsprechende Szenarien selbst generieren ließen. Die aus Fragebogenstudien resultierenden Informationen bleiben jedoch auf die vorgegebenen Merkmale beschränkt. Dadurch ist es möglich, dass weitere Indikatoren, die für die Glaubhaftigkeitsattribution verwendet werden, nicht erfasst werden.

Daher werden Versuchspersonen bei Beurteilungsstudien instruiert zuvor abgegebene Glaubhaftigkeitsurteile frei zu begründen (Anderson, DePaulo, Ansfield, Tickle & Green, 1999; Feeley & Young, 2000; Freedman, Adam, Davey & Koegel, 1996, Experiment 3; Granhag & Strömwall, 2000, 2001; Hartwig, Granhag, Strömwall & Andersson, 2004; Landström, Granhag & Hartwig, 2005; Mann et al., 2004; Strömwall, Granhag & Landström, 2007; Reinhard, Burghardt, Sporer &

Bursch, 2002). Doch auch hier bleibt unklar, ob die angeführten Urteilsbegründungen tatsächlich entscheidungsleitend waren.

Dies lässt sich unter Verzicht auf introspektive Daten über eine Brunswiksche Linsenmodellanalyse überprüfen (Fiedler, 1989a, Studie 1, 1989b; Köhnken, 1990; Kraut, 1978, Experiment 1; Reinhard, Burghardt, Sporer & Bursch, 2002; Sporer & Küpper, 1995). Dabei werden die im Untersuchungsmaterial vorliegenden Aussagemerkmale einerseits mit dem subjektiven Glaubhaftigkeitsurteil, andererseits mit dem objektiven Wahrheitsstatus in Beziehung gesetzt. So zeigen die Ergebnisse nicht nur, in welchem Ausmaß beobachtbare Aussagemerkmale bei der Glaubhaftigkeitsattribution berücksichtigt werden, sondern auch inwieweit sie mit dem objektiven Wahrheitsstatus korrespondieren.

Schließlich liefern Untersuchungen zu Täuschungsstrategien indirekte Hinweise darauf, welche Merkmale Personen mit wahren oder erfundenen Aussagen assoziieren (Fiedler, 1989a, Studie 2; Niehaus, Krause & Schmidke, 2005).

Forschungsstand zu subjektiven Annahmen

Inhaltliche Aussagemerkmale werden im Rahmen von Beurteilungsstudien häufig als Urteilsbegründungen angeführt (z.B. Anderson et al., 1999; Granhag & Strömwall, 2001; Hartwig et al., 2004; Landström et al., 2005; Strömwall et al., 2007). Im Vergleich zu nonverbalen Indikatoren wurden subjektive Glaubhaftigkeitsannahmen zu inhaltlichen Aussageaspekten jedoch seltener und weniger umfassend untersucht (vgl. die zusammenfassenden Übersichten von Strömwall et al., 2004; Vrij 2008). Zudem ist der bisherige Forschungsstand sehr unübersichtlich, da Anzahl und Art der untersuchten Indikatoren stark variieren. Nur selten wurde die Terminologie etablierter objektiver Glaubhaftigkeitsmerkmale eingehalten (Akehurst et al., 1996; Reinhard et al., 2002; Taylor & Vrij, 2000; Vrij et al., 2001, 2006), um Indikatoren zur Beurteilung vorzugeben oder Kategorien für

die Analyse freier Urteilsbegründungen abzuleiten. Doch selbst wenn sich für einzelne Merkmale einheitliche Begrifflichkeiten finden, wurden sie oftmals unterschiedlich operationalisiert. Beispielsweise bezogen einige Autoren die logische Konsistenz auf den Vergleich wiederholter Aussagen (Granhag et al., 2004, 2005; Granhag & Strömwall, 2000, 2001; Hartwig et al., 2004; Kraut, 1978; Strömwall & Granhag, 2003). Andere Autoren wiederum bezogen sich auf die logische Konsistenz innerhalb einer Aussage oder führten nicht weiter aus, wie dieses Aussagemerkmal aufzufassen sei (Akehurst et al., 1996; Reinhard et al., 2002; Taylor & Vrij, 2000; Vrij et al., 2001, 2006). Aufgrund der Unterschiede in den verwendeten Begrifflichkeiten, Forschungsparadigmen und Auswertungsmethoden lässt sich der bisherige Forschungsstand zu subjektiven Täuschungsindikatoren nur qualitativ zusammenfassen. In Tabelle 2.1 ist aufgeführt welche Aussagemerkmale eher mit wahren bzw. erfundenen Aussagen assoziiert werden. Für die tabellarische Übersicht wurde die Terminologie der CBCA-Kriterien verwendet und durch das RÜ-Kriterium kognitive Operationen sowie durch das oftmals untersuchte inhaltliche Merkmal der Plausibilität ergänzt. Um den bisherigen Forschungsstand zusammenfassend zu interpretieren, wurden jedoch auch Studien miteinbezogen, die davon abweichende Bezeichnungen verwendeten. Die wörtlich übersetzten Original-Bezeichnungen finden sich bei den folgenden Ausführungen zu den Befunden einzelner Studien.

Bezüglich der globalen Merkmale zeichnen sich weitestgehend einheitliche Forschungsbefunde ab. Auf die Frage hin, woran man erkennt, dass Personen lügen, wurde besonders häufig auf inkohärente Darstellungen und verbale Inkonsistenzen verwiesen (Global Deception Research Team, 2006, Studie 1). Ebenso schätzten Polizeibeamte, die befragt wurden, anhand welcher Merkmale sie die Glaubhaftigkeit von Aussagen zu inkriminierten Sexualdelikten beurteilen, inhaltliche Inkonsistenzen und einen Mangel an Plausibilität als Lügenindikatoren ein (Greuel, 1992). Entsprechend zeigten auch Fragebogenuntersuchungen, dass logische Inkonsistenzen oft als Lügenindikatoren gewertet wurden.

Tabelle 2.1

Zusammenfassende Übersicht der bisherigen Forschungsbefunde zu subjektiven Annahmen hinsichtlich inhaltlicher Täuschungskorrelate

Inhaltliche Indikatoren	Subjektive Annahmen
Logische Konsistenz	+
Plausibilität	+
Widersprüche	-
Raum-Zeitliche Verknüpfungen	+/0/-
Unstrukturierte Darstellung	-
Quantitativer Detailreichtum	+/-
Interaktionschilderungen	+
Wiedergabe von Gesprächen	+
Schilderung von Komplikationen im Handlungsverlauf	-
Schilderung ausgefallener Einzelheiten	-
Schilderung nebensächlicher Einzelheiten	-
Schilderung eigener psychischer Vorgänge	+/0/-
Schilderung psychischer Vorgänge des Täters/ anderer	+/0/-
Spontane Verbesserungen der eigenen Aussage	-/0
Eingeständnis von Erinnerungslücken	-/0
Einwände gegen die Richtigkeit der eigenen Aussage	+/0/-
Selbstbelastungen	+/0
Kognitive Operationen	+/0

Anm. + indiziert eine subjektive Einschätzung als Wahrheitsindikator, - als Lügenindikator; 0 indiziert, dass hinsichtlich dieses Merkmals keine Unterschiede zwischen wahren und erfundenen Aussagen erwartet werden.

Dies gilt sowohl für Studien, bei denen die Inkonsistenz zwischen mehreren Aussagen betrachtet wurde (Granhag et al., 2004, 2005; Strömwall & Granhag, 2003) als auch für Studien, die nicht spezifizierten wie das Merkmal aufzufassen sei (Akehurst et al., 1996; Global Deception Research Team, 2006, Studie 2; Vrij et al., 2001, 2006). Zudem wurde meist angenommen, dass erlogene Aussagen weniger plausibel seien und mehr Widersprüche aufweisen würden als wahre Aussagen (Granhag & Strömwall, 2000; Köhnken, 1990, Lakhani & Taylor, 2003; Vrij et al., 2006). Obwohl häufig angenommen wurde, dass Lügen unstrukturierter vorgetragen werden (Akehurst et al., 1996; Vrij et al., 2001, 2006), wurden spontane Aussagen für glaubhafter gehalten als Aussagen, die anscheinend vorbereitet und gedanklich wiederholt wurden (Granhag & Strömwall, 2000).

Ein hoher Detaillierungsgrad wurde ebenfalls in den meisten Studien mit wahren Aussagen assoziiert (Strömwall & Granhag, 2003; Vrij et al., 2001, 2006). Entsprechend zeigten auch Untersuchungen zu Täuschungsstrategien, dass Personen gezielt Details in eine Falschaussage integrieren würden, um überzeugend zu wirken (Niehaus et al., 2005) bzw. auch tatsächlich entsprechend handelten (Fiedler, 1989a, Studie 2). Im Gegensatz dazu berichteten Granhag und Strömwall (2000), dass Personen sich uneinig waren, ob ein hoher Detaillierungsgrad mit wahren oder erfundenen Aussagen zu assoziieren sei. Zudem nahmen Studierende, Gefängnisinsassen und -bedienstete in zwei Fragebogenuntersuchungen an, dass erfundene Aussagen mehr Details beinhalten würden als wahre (Granhag et al., 2004, 2005). Die Autoren argumentierten, dass dieser ungewöhnliche Befund auf die Berücksichtigung von Täuschungsstrategien zurückzuführen sei. Es wurde nämlich ebenso erwartet, dass Lügner sich stärker auf den Inhalt ihrer Aussage vorbereiten würden als Personen, die wahr aussagen (Granhag et al., 2005), und dass es erfolgsförderlich sei Falschaussagen zu planen (Granhag et al., 2004). Entsprechend ließe sich die Annahme detailreicher Falschaussagen dadurch erklären, dass Lügnern unterstellt wurde sich auf ihre Aussage vorzubereiten,

indem sie gezielt Details integrieren. Zudem fragten Granhag et al. (2004, 2005) allgemein nach der Menge an Details, ohne deren Qualität zu spezifizieren.

Andere Untersuchungen zeigten jedoch, dass spezifische Details eher mit erfundenen Aussagen assoziiert wurden. Dies gilt für einige Merkmale, die den Kategorien der speziellen Inhalte und inhaltlichen Besonderheiten zugeordnet sind. Beispielsweise wurde erwartet, dass ausgefallene und nebensächliche Einzelheiten häufiger in erfundenen als in wahren Aussagen vorzufinden seien (Akehurst et al., 1996; Fiedler, 1989a, 1989b; Forrest et al., 2004; Taylor & Vrij, 2000; Zuckerman, Koestner & Driver, 1981). Zudem wurde eine Zunahme hinsichtlich der Schilderung von Komplikationen im Handlungsverlauf bei erfundenen im Vergleich zu wahren Aussagen erwartet (Akehurst et al., 1996; Taylor & Vrij, 2000). Im Gegensatz dazu wurden Schilderungen von Interaktionen und Gesprächen eher mit wahren Aussagen assoziiert (Akehurst et al., 1996; Vrij et al., 2001).

In einzelnen Untersuchungen wurde zudem angenommen, erlogene Aussagen seien länger als wahre (Global Deception Research Team, 2006; Lakhani & Taylor, 2003). Sowohl die Erwartung, dass Falschaussagen gezielt mit erfundenen Details angereichert würden, als auch dass spezifische Details zunehmen würden, könnten diesen Befund erklären. Schließlich gibt es Hinweise auf die subjektive Annahme, dass erfundene Aussagen mehr kognitive Anstrengung erfordern würden als wahre (Granhag et al., 2004, 2005).

Für die Aussagemerkmale Raum-zeitliche Verknüpfungen, Schilderungen eigener psychischer Vorgänge und denen des Täters wurden weniger einheitliche Befunde berichtet. Nach Akehurst et al. (1996) glauben Personen, dass Raum-zeitliche Verknüpfungen zunehmen, wenn sie selbst lügen, jedoch abnehmen, wenn andere lügen. Vrij et al. (2001) wiederum fanden am häufigsten die subjektive Annahme, dass sich wahre und erfundene Aussagen nicht hinsichtlich Raum-zeitlicher Verknüpfungen unterscheiden würden. Hintergrundinformationen zum geschilderten Ereignis wurden jedoch nach Taylor und Vrij (2000) eher bei

erfundenen als bei wahren Aussagen erwartet. Hinsichtlich der Schilderung eigener psychischer Vorgänge berichteten Taylor und Vrij (2000), dass keine signifikanten Unterschiede zwischen wahren und erfundenen Aussagen erwartet wurden. Die Schilderung eigener Gefühle wurde jedoch in anderen Studien als Lügenindikator (Akehurst et al., 1996) oder Wahrheitsindikator (Vrij et al., 2006) aufgefasst. Auch Selbstbezügen wurde entweder keinerlei Differenzierungskraft zugesprochen (Forrest et al., 2004) oder sie wurden eher bei erfundenen Aussagen erwartet (Zuckerman, Koestner & Driver, 1981). Hinsichtlich der Schilderung psychischer Vorgänge des Täters wurden nach Vrij et al. (2001) keine Unterschiede zwischen erfundenen und wahren Aussagen erwartet. Akehurst et al. (1996) berichteten widersprüchliche subjektive Annahmen für die Schilderung von Gefühlen anderer. Für eigene Aussagen wurde eine Zunahme, für die Aussagen anderer eine Abnahme dieses Merkmals infolge des Lügens erwartet. Täuschungsstrategisch wurde es als sinnvoll erachtet räumliche Verknüpfungen sowie Schilderungen eigener psychische Vorgänge in Falschaussagen zu integrieren, während zeitliche Verknüpfungen und Verweise auf fremdpsychisches Erleben als irrelevant erachtet wurden (Niehaus et al., 2005).

Die Validität einer Reihe von Glaubhaftigkeitsmerkmalen wird auf stereotype Täuschungsannahmen zurückgeführt (z.B. DePaulo et al., 2003). Demnach sollten spontane Verbesserungen der eigenen Aussage, Eingeständnisse von Erinnerungslücken, Einwände gegen die Richtigkeit der eigenen Aussage und Selbstbelastungen als Lügenindikatoren aufgefasst werden. Allerdings ist der Forschungsstand diesbezüglich wenig einheitlich.

Bei Beurteilungsstudien wurden Eingeständnisse von Erinnerungslücken, Fehlern und sozial Unerwünschtem nur selten als freie Urteilsbegründungen angeführt (Reinhard et al., 2002). Dennoch wurden in Fragebogenuntersuchungen vereinzelt entsprechende Merkmale zur Bewertung vorgegeben. Demnach wurde bezüglich der Eingeständnisse von Erinnerungslücken entweder angenommen, dass diese nicht zwischen wahren und erlogenen Aussagen differenzieren

würden (Lakhani & Taylor, 2003; Vrij et al., 2001), oder dass sie häufiger bei erlogenen Aussagen vorzufinden seien (Akehurst et al., 1996; Taylor & Vrij, 2000; Vrij et al., 2006). Spontane Verbesserungen der eigenen Aussage wurden eher mit erlogenen Aussagen assoziiert (Akehurst et al., 1996; Taylor & Vrij, 2000). Jedoch fanden Vrij et al. (2001) am häufigsten die subjektive Annahme, dass spontane Verbesserungen der eigenen Aussage bei wahren und falschen Aussagen gleichermaßen häufig vorzufinden seien, d.h. dass dieses Merkmal nicht zwischen wahren und falschen Aussagen differenzieren würde. In einigen Untersuchungen wurden Glaubhaftigkeitsurteile damit begründet, dass der Sender entweder wenig, oder umgekehrt sehr sicher und überzeugt gewirkt hat (Granhag & Strömwall, 2000; Hartwig et al., 2004). Allerdings schien keine Einigkeit darüber zu bestehen, ob eine sichere Darstellungsweise mit wahren oder falschen Aussagen zu assoziieren sei (Granhag & Strömwall, 2000). Entsprechend ergaben sich auch bei Fragebogenuntersuchungen uneinheitliche Befunde. Während bei Akehurst et al. (1996) angenommen wurde, Lügner würden eher als wahr aussagende Personen Einwände gegen die Richtigkeit der eigenen Aussage vorbringen, zeigte sich bei Taylor und Vrij (2000) die gegenteilige subjektive Annahme. In einer Studie von Vrij et al. (2001) wiederum assoziierten die meisten Probanden dieses Merkmal weder mit wahren noch mit falschen Aussagen.

Auch für das Ausmaß an Selbstbelastungen ist die Befundlage eingeschränkt und uneindeutig. Entweder wurde angenommen, dass sich wahre und erlogene Aussagen diesbezüglich nicht unterscheiden würden (Akehurst et al., 1996), oder dass Lügner weniger Selbstbelastungen äußern würden (Taylor & Vrij, 2000). Täuschungsstrategisch wurde es im Kontext von Sexualdelikten als sinnvoll erachtet, Einwände gegen die eigene Aussage, spontane Verbesserungen und Selbstbelastungen zu vermeiden, während Eingeständnisse von Erinnerungslücken als wenig relevant erachtet wurden (Niehaus et al., 2005).

Subjektive Annahmen wurden für die inhaltlichen Aussagemerkmale phänomengemäße Schilderung unverstandener Handlungselemente, indirekt

handlungsbezogene Schilderungen, Entlastung des Angeschuldigten und deliktspezifische Aussageelemente bislang nicht erfasst. Die Befunde von Niehaus et al. (2005) verwiesen jedoch darauf, dass Personen bei Falschaussagen zu Sexualdelikten Entlastungen des Angeschuldigten vermeiden würden.

Schließlich liegen verschiedene Untersuchungen zu subjektiven Annahmen hinsichtlich des RÜ-Merkmals kognitive Operationen vor. In einer Fragebogenstudie wurde dieses inhaltliche Aussagemerkmal als Wahrheitsindikator aufgefasst (Vrij et al., 2006). Analysen anhand des Brunswikschen Linsenmodells wiesen entweder einen positiven (Reinhard et al., 2002) oder keinen Zusammenhang zwischen kognitiven Operationen und dem subjektiven Glaubhaftigkeitsurteil nach (Sporer & Küpper, 1995).

Insgesamt ist der Forschungsstand für einige Merkmale relativ einheitlich, während er für andere Merkmale, unter anderem auch für stereotypkonträre, als widersprüchlich zu bewerten ist. Unterschiede im inhaltlichen Verständnis der Aussagemerkmale tragen möglicherweise zu den divergierenden Forschungsbefunden bei, können diese aber nicht vollständig erklären. Daher erscheint es sinnvoll zu kontrollieren, auf welche Art von Täuschungssituation die Probanden ihre subjektiven Annahmen beziehen.

Varianten von Fragebogenstudien

Das Global Deception Research Team (2006) führte eine weltweite Befragung zu subjektiven Annahmen von Täuschungskorrelaten durch. Auf die Frage hin, woran man erkennt, ob jemand lügt, wurde nach Ansicht der Autoren überraschend selten (0.33%) auf situationale Faktoren, z.B. spezifische Anreize zu täuschen, verwiesen. Dass Personen entsprechende Aspekte nicht von sich aus benennen, bedeutet allerdings nicht notwendigerweise, dass sie für situationsspezifische Erfordernisse nicht sensitiv sind und ihre subjektiven Annahmen nicht danach ausrichten. Um dies zu prüfen ist es notwendig, situative

Aspekte durch experimentelle Vorgaben gezielt zu variieren. Dieses Vorgehen wurde von Breuer et al. (2005) sowie von Taylor und Hick (2007) gewählt, um subjektive Annahmen zu non- und paraverbalen Merkmalen zu erfassen. Für subjektive Annahmen zu inhaltlichen Aussagemerkmalen liegen unseres Wissens derzeit nur drei Fragebogenuntersuchungen vor (Lakhani & Taylor, 2003; Taylor & Hill-Davies, 2004; Taylor & Vrij, 2000), die potenzielle Moderatoren durch die Vorgabe fiktiver Szenarien variierten.

Taylor und Hill-Davies (2004) interessierten sich vor allem für Effekte des Alters der lügenden Person auf subjektive Annahmen zu Täuschungsindikatoren. Daher variierten sie über die Vorgabe fiktiver Szenarien im Rahmen eines Within-Subjects-Designs das Alter der lügenden Person und schilderten jeweils einen altersgemäßen Gegenstand der Lüge. So wurde beispielsweise ausgeführt, dass ein Kind im Vorschulalter lügen würde, weil es trotz eines Verbots einen Keks genommen hat, während ein jugendliches Mädchen lügen würde, weil sie eine Party gefeiert hat, als ihre Eltern außer Haus waren. Neben 22 nonverbalen und paraverbalen Merkmalen wurden die beiden inhaltlichen Merkmale spontane Korrekturen und Beschuldigung anderer zur Bewertung vorgegeben. Hinsichtlich der subjektiven Täuschungsannahmen zu spontanen Korrekturen zeigten sich keine Unterschiede zwischen den Szenarien. Es wurde jedoch angenommen, dass jüngere Kinder eher andere beschuldigen würden als ältere. Effekte des Alters und des Gegenstands der Lüge waren in dieser Studie jedoch konfundiert.

Die Untersuchungen von Lakhani und Taylor (2003) sowie von Taylor und Vrij (2000) konzentrierten sich hingegen auf situative Aspekte. Taylor und Vrij (2000) variierten im Rahmen eines Between-Subjects-Designs die Relevanz der Täuschungssituation (hoch/mittel/gering) sowie die kognitive Komplexität (leicht/schwierig) einer Falschaussage. Während sich kein Einfluss der kognitiven Komplexität auf die subjektiven Täuschungsannahmen zeigte, wurden drei von zwölf inhaltlichen Aussagemerkmalen situationsspezifisch bewertet. Neben dem Ausmaß an Hintergrundinformationen galt dies für die beiden stereotypkonträren

Merkmale Eingeständnisse von Erinnerungslücken und Selbstbelastungen. Die Autoren führten interpretierend aus, dass insbesondere in heiklen Situationen von Lügnern Verhaltensweisen erwartet werden, die auch ein Verkäufer zeigen würde, um die Glaubhaftigkeit seiner Aussage zu erhöhen.

Lakhani und Taylor (2003) wählten ein Within-Subjects-Design, um den Einfluss der Täuschungsrelevanz (gering/hoch) auf subjektive Täuschungsindikatoren zu überprüfen. Dazu wurden ebenfalls fiktive Szenarien vorgegeben, die eine alltägliche und eine rechtlich relevante Lügensituation darstellten. Von sechs untersuchten inhaltlichen Aussagemerkmalen wurden zwei situationsspezifisch bewertet. Die Versuchspersonen nahmen an, dass Lügen in rechtlich relevanten Situationen (hohe Relevanz) widersprüchlicher seien und eine geringere logische Konsistenz aufweisen würden als in Alltagssituationen. Hingegen zeigten sich keine Unterschiede zwischen den Täuschungssituationen bezüglich der Anzahl an Klischees, der Plausibilität der Aussagen, des Detaillierungsgrades der Antworten und der Tendenz Erinnerungslücken einzugestehen. Weitere Unterschiede zwischen den Szenarien zeigten sich in den subjektiven Annahmen zu nonverbalen Indikatoren. Daher schlussfolgerten die Autoren, dass die Relevanz einer Täuschungssituation durchaus die subjektiven Annahmen zu Täuschungskorrelaten beeinflusst.

Ziele und Hypothesen

Im Rahmen der vorliegenden Untersuchung wurden ebenfalls subjektive Annahmen vor dem Hintergrund fiktiver Szenarien erfasst. Im Gegensatz zu der Untersuchung von Taylor und Hill-Davies (2004) standen dabei keine personengebundenen Merkmale sondern situative Aspekte im Vordergrund. Zudem wurden im Vergleich zu den Untersuchungen von Lakhani und Taylor (2003) sowie Taylor und Vrij (2000) wesentlich mehr inhaltliche Aussagemerkmale zur Beurteilung vorgelegt. Lakhani und Taylor untersuchten sechs, Taylor und Vrij zwölf inhaltliche Merkmale, wobei sie für vier bzw. elf

Merkmale die CBCA-Terminologie verwendeten. Hingegen wurden im Rahmen der vorliegenden Untersuchung 52 inhaltliche Aussagemerkmale zur Beurteilung vorgegeben, die aus den ARJS-Items abgeleitet wurden. Deren Originalbezeichnungen wurden dabei beibehalten. Die 52 Einzelitems wurden anschließend zu den 13 ARJS-Skalen zusammengefasst. Zudem wurde im Rahmen eines 3 x 2 Between-Subjects-Designs nicht nur die Relevanz der Täuschungssituation (gering/mittel/hoch) variiert, sondern auch das Ausmaß der vermeintlichen Vorbereitungszeit (kurz/lang). Dies erschien nicht nur vor dem Hintergrund einer höheren ökologischen Validität sinnvoll, sondern auch weil gezeigt wurde, dass die Urteilsrichtigkeit bei vorbereiteten Aussagen geringer ist als bei unvorbereiteten (Bond & DePaulo, 2006). Das vollständige Untersuchungsdesign ist Tabelle 2.2 zu entnehmen.

Tabelle 2.2

Aufteilung der Stichprobe zu den verschiedenen Untersuchungsbedingungen

	Vorbereitungszeit	
Situation	Kurz	Lang
Verspätung	n = 40	n = 40
Affäre	n = 40	n = 40
Totschlag	n = 40	n = 40
Sexualdelikt	---a	n = 40
Kontrollgruppe	n = 40	

Anm. N = 340, ---a Eine Bedingung mit kurzer Vorbereitung wurde für das Szenario zum Sexualdelikt nicht realisiert.

Das Szenario geringer Relevanz bezog sich auf eine Alltagslüge wegen einer Verspätung, das Szenario mittlerer Relevanz auf eine Lüge wegen einer außereheliche Affäre und das Szenario hoher Relevanz auf eine strafrechtlich relevante Falschaussage wegen eines Totschlagsdeliktes. Für jede Täuschungssituation wurde expliziert, dass die lügende Person sich entweder nur

kurz oder lange auf ihre Aussage vorbereitet hat. Eine weitere Experimentalgruppe erhielt als fiktives Szenario eine lang vorbereitete Falschaussage hinsichtlich eines Sexualdeliktes. Zudem wurde eine Kontrollgruppe realisiert, der kein fiktives Täuschungsszenario vorgegeben wurde.

Alle Versuchspersonen gaben an, ob sie die 52 ARJS-Items eher mit wahren oder erfundenen Aussagen assoziieren. Die Beurteilungen der einzelnen Items lassen sich gemäß der Struktur der ARJS 13 Skalen zuordnen. Die Validität einiger CBCA-Merkmale wurde darauf zurückgeführt, dass sie stereotypen Vorstellungen von Täuschung entsprechen und deswegen bei Falschaussagen gezielt vermieden werden (Ruby & Brigham, 1998; Steller & Köhnken, 1989). Die Skala Fehler und sozial Unerwünschtes der ARJS umfasst die Merkmale „das Ereignis sicher und überzeugt darstellen“, „Fehler oder Ungenauigkeiten spontan verbessern“, „Erinnerungslücken hinsichtlich einiger Details zugeben“ und „unschmeichelhafte Handlungen aufgrund persönlicher Schwächen, Irrtümer und Fehler zugeben“. Die ARJS-Merkmale sind demnach anders operationalisiert als die stereotypkonträren CBCA-Merkmale, dennoch sind gewisse Ähnlichkeiten festzustellen. So nimmt auch Sporer (1996/1998/2004) an, dass diese ARJS-Merkmale aufgrund von Selbstdarstellungsbemühungen bei erfundenen Aussagen gezielt vermieden werden. Daher wurde für die vorliegende Untersuchung postuliert, dass die Merkmale der ARJS-Skala Fehler und sozial Unerwünschtes subjektiv mit Lügen assoziiert werden (Hypothese 1a). Zudem sollten dem bisherigen Forschungsstand zu subjektiven Annahmen über Täuschungskorrelate entsprechend (Akehurst et al., 1996; Fiedler, 1989a, Studie 1, 1989b; Taylor & Vrij, 2000; Vrij et al., 2006) die unter der ARJS-Skala Komplikationen und ungewöhnliche Details zusammengefassten Merkmale als Lügenindikatoren beurteilt werden (Hypothese 1b). Weitere inhaltliche Glaubhaftigkeitsmerkmale, beispielsweise der quantitative Detailreichtum, das Ausmaß an Interaktionsschilderungen und die logische Konsistenz von Aussagen, wurden oftmals auch subjektiv mit wahren Aussagen assoziiert. Da die

ARJS ebenso wie die CBCA objektiv als Glaubhaftigkeitsmerkmale aufzufassen sind, wird angenommen, dass die meisten Skalen auch subjektiv mit wahren Aussagen assoziiert werden. Andererseits könnten deutliche Diskrepanzen zwischen objektiven und subjektiven Indikatoren dazu beitragen, die Validität einzelner ARJS-Skalen zu erklären.

Zudem wurde ein Effekt der variierten Täuschungssituation auf die subjektiven Annahmen erwartet. Mit zunehmender Relevanz der Situation sollten sowohl die negativen Konsequenzen bei der Überführung einer Lüge als auch das kritische Hinterfragen der Aussage zunehmen. Demnach wäre die Motivation zu überzeugen und die Schwierigkeit zu täuschen in rechtlich relevanten Situationen höher als in Alltagssituationen. Wenn Personen diese situationalen Aspekte berücksichtigen, sollte sich ein Haupteffekt der systematisch variierten Relevanz der Täuschungssituation (gering/mittel/hoch) auf die subjektiven Annahmen zeigen. Aufgrund der erhöhten Schwierigkeit zu täuschen sollten die subjektiven Annahmen für Situationen hoher Relevanz stärker ausgeprägt sein als für Situationen geringer Relevanz (Hypothese 2).

Ebenso ist anzunehmen, dass die experimentell induzierte Variation der Vorbereitungszeit die subjektiven Annahmen beeinflusst. In einer Studie von Granhag et al. (2004) nahmen Personen an, Lügner würden von einer intensiven Planung ihrer Aussage profitieren. Daher wurde für die vorliegende Untersuchung postuliert, dass die subjektiven Annahmen zu inhaltlichen Täuschungskorrelaten geringer ausfallen, wenn die lügende Person Gelegenheit hatte ihre Aussage vorzubereiten, als wenn sie sich überraschend äußern muss (Hypothese 3).

Des Weiteren sollen potenzielle Wechselwirkungseffekte zwischen der Täuschungssituation und der Vorbereitungszeit klären, wann Unterschiede zwischen wahren und erlogenen Aussagen erwartet werden. Insbesondere die subjektiven Annahmen zu der ARJS-Skala Details, bei der sowohl qualitative als auch quantitative Aspekte von Details zu berücksichtigen sind, könnten dadurch ausdifferenziert werden. Dies erscheint unter anderem daher interessant, weil für

das CBCA-Merkmal des quantitativen Detailreichtums widersprüchliche subjektive Annahmen festzustellen waren (z.B. Granhag & Strömwall, 2000).

Die meisten Untersuchungen zeigten, dass Details subjektiv mit wahren Aussagen assoziiert wurden. Entsprechend wurde es auch täuschungsstrategisch als sinnvoll erachtet, Aussagen inhaltlich anzureichern (Niehaus et al., 2005). Möglicherweise wird aber ebenso angenommen, dass die Umsetzung solcher Täuschungsstrategien ein gewisses Ausmaß an Planung voraussetzt. Dies würde sich in der subjektiven Annahme widerspiegeln, dass Lügner in strafrechtlich relevanten Situationen nicht zu detailreichen Darstellungen in der Lage sind, wenn sie keine Gelegenheit hatten ihre Aussage vorzubereiten. Der Verzicht auf Details könnte die vermeintliche Gefahr verringern, sich in Widersprüche zu verwickeln bzw. falsifizierbare Angaben zu machen. Entsprechend sollten für Situationen hoher Relevanz erfundene Aussagen mit einem geringeren Detaillierungsgrad assoziiert werden als wahre, wenn sie unvorbereitet vorgebracht werden (Hypothese 4a). Besteht jedoch Gelegenheit sich vorzubereiten, wird möglicherweise angenommen, dass Lügner diese gezielt nutzen, um Details zu erfinden. Infolgedessen würde für solche Situationen die subjektive Annahme bezüglich eines unterschiedlichen Detaillierungsgrades wahrer und erfundener Aussagen weniger deutlich ausfallen (Hypothese 4b). In Bezug auf Alltagssituationen haben die Probanden möglicherweise selbst erfahren, dass Aussagen selten angezweifelt werden und es als weniger wichtig erachtet wird zu überzeugen. Demnach sind auch weniger täuschungsstrategische Bemühungen, wie das Erfinden falscher Details, zu erwarten. Daher wurde postuliert, dass sich für Situationen geringer Relevanz unabhängig von der verfügbaren Vorbereitungszeit die subjektive Annahme zeigt, erfundene Aussagen seien weniger detailliert als wahre (Hypothese 4c).

Zudem sollte überprüft werden, ob Personen bei der Beurteilung von Indikatoren eher an Täuschungssituationen mit hoher Relevanz oder eher an Alltagssituationen mit geringer Relevanz denken. Lakhani und Taylor (2003)

führten aus, dass Personen möglicherweise auf der Grundlage von Situationen mit hoher Relevanz stereotype Vorstellungen über Lügen ausbilden. Auch Zuckerman, Koestner und Driver (1981) spekulierten, dass Personen bei der Befragung zu subjektiven Täuschungsannahmen möglicherweise an hoch motivierte Lügner denken, die sich mittelmäßig auf ihre Falschaussage vorbereitet haben. Daher wurde eine Kontrollgruppe, die keine Situationsvorgabe erhielt, zu ihren subjektiven Annahmen befragt. Den Überlegungen von Lakhani und Taylor sowie Zuckerman, Koestner und Driver entsprechend sollte die Kontrollgruppe gegenüber den Gruppen mit vorgegebener Alltagssituation stärkere Verhaltensänderungen erwarten und sich in ihren subjektive Annahmen weniger von den Gruppen, die ihre Angaben auf eine Täuschungssituation mit hoher Relevanz beziehen, unterscheiden (Hypothese 5).

Schließlich sollte untersucht werden, ob es spezifische subjektive Annahmen für Falschaussagen bei Sexualdelikten gibt. Dazu wurden die subjektiven Annahmen zweier Bedingungen verglichen, in denen eine hohe Relevanz einerseits über ein Totschlagsdelikt, andererseits über ein Sexualdelikt, operationalisiert wurde. Unterscheiden sich die subjektiven Annahmen dieser beiden Bedingungen, bestätigt dies die Hypothese der Situationsspezifität von Sexualdelikten. Niehaus et al. (2005) verwiesen auf den besonders „kompromittierenden Charakter“ (S. 179) der Inschutznahme des Angeschuldigten und Selbstbelastungen im Kontext von Sexualdelikten. Bei den ARJS wird unter anderem die Tendenz, negative Äußerungen über sich selbst abzugeben, durch die Skala Fehler und sozial Unerwünschtes erfasst. Analog der Argumentation von Niehaus et al. wurde postuliert, dass diese Skala in der Bedingung zum Sexualdelikt deutlicher mit erfundenen Aussagen assoziiert wird als in der Bedingung zum Totschlagsdelikt (Hypothese 6).

Methode

Versuchspersonen

Insgesamt bearbeiteten N = 320 Versuchspersonen einen Fragebogen zu subjektiven Annahmen über Täuschungskorrelate. Dabei wurden gleich viele Männer und Frauen befragt, die überwiegend studierten (85.9%), wobei auch einige noch Schüler (5.0%) oder bereits berufstätig waren (9.1%). Die Versuchspersonen waren zwischen 14 und 72 Jahre alt, mit einem Altersdurchschnitt von 23.85 Jahren (SD = 6.65, Mdn = 22).

Untersuchungsdesign und Stimulusmaterial

Die Versuchspersonen wurden wie in Tabelle 2.2 dargestellt zufällig einer von sieben Experimentalbedingungen oder einer Kontrollgruppe zugewiesen. Die Experimentalgruppen unterschieden sich in der Vorgabe eines fiktiven Szenarios, auf das sie ihre Angaben beziehen sollten. Die verwendeten Szenarien sind im Anhang 2a dokumentiert. Im Rahmen eines 3 x 2 Between-Subjects-Designs wurden die Relevanz der Täuschungssituation (gering/mittel/hoch) sowie das Ausmaß der verfügbaren Vorbereitungszeit (kurz/lang) systematisch variiert. Zusätzlich wurde neben einer Kontrollgruppe ohne spezifische Situationsvorgabe eine weitere Experimentalgruppe realisiert (s.u.). In der Täuschungssituation geringer Relevanz wurde gelogen, um eine Verspätung zu rechtfertigen. Eine mittlere Relevanz wurde dadurch operationalisiert, dass eine Person ihren Ehepartner belügt, um eine außereheliche Affäre zu verheimlichen. Für die Täuschungssituation hoher Relevanz wurde ein rechtlich relevantes Szenario gewählt. Dabei sagte eine Person gegenüber der Polizei falsch aus, um sich der Verantwortung für einen Totschlag im Affekt zu entziehen. In allen Szenarien wurde entweder darauf hingewiesen, dass die lügende Person sich überraschend äußern musste (kurze Vorbereitung) oder dass sie ausreichend Zeit hatte, ihre Aussage zu planen (lange Vorbereitung). Zudem wurde in den daraus resultierenden sechs Experimentalbedingungen die aussagende Person in

jeweils der Hälfte der Szenarien als weiblich (Frau K.) bzw. als männlich (Herr K.) beschrieben. Dadurch wurden mögliche Unterschiede in Abhängigkeit von dem Geschlecht der Stimulusperson ausbalanciert.

Eine weitere Experimentalbedingung erhielt als fiktives Szenario eine Falschaussage bezüglich eines Sexualdeliktes, das von Niehaus et al. (2005) adaptiert wurde. Dabei waren nur geringfügige Veränderungen notwendig, um die Vergleichbarkeit dieses Szenarios mit den anderen, in der vorliegenden Studie verwendeten, Szenarien zu gewährleisten. Niehaus et al. instruierten die Versuchspersonen sich vorzustellen, sie würden selbst falsch aussagen. Im Gegensatz dazu wurde die lügende Person in der vorliegenden Untersuchung erneut als Frau K. bezeichnet. Das Szenario zum Sexualdelikt beschrieb, dass Frau K. sich nach gründlicher Vorbereitung der Polizei gegenüber als Opfer eines sexuellen Mißbrauchs darstellt, obwohl tatsächlich ihre Freundin dieses Opfer war. Für diese Täuschungssituation wurde weder eine Bedingung mit kurzer Vorbereitungszeit noch mit einer männlichen Stimulusperson realisiert, da beides unrealistisch erschien.

Fragebogen

Alle Probanden bearbeiteten den gleichen Fragebogen zu subjektiven Annahmen über Täuschungskorrelate. Dieser umfasste insgesamt 73 Items, von denen 52 den ARJS-Items entsprachen und 14 sich auf die allgemeine Eindrucksbildung bezogen. Beispielsweise wurde gemäß der ARJS gefragt: “Im Vergleich zu einer wahren Aussage werden bei einer erlogenen Aussage (weniger/mehr) unplausible Dinge erwähnt”. Zur allgemeinen Eindrucksbildung wurde beispielsweise das Item “Im Vergleich zu einer wahren Aussage wirkt die Person bei einer erlogenen Aussage eher (nervös/ruhig)“ zur Beurteilung vorgegeben. Die Probanden beurteilten jedes Item anhand 7-stufiger Skalen von –3 bis +3, deren Endpunkte quantitativ (z.B. seltener–-häufiger) oder qualitativ (z.B. vage–-klar) verankert waren. Negative Werte verwiesen auf eine vermeintliche

Abnahme, positive auf eine vermeintliche Zunahme der Quantität oder Qualität des entsprechenden Merkmals beim Lügen. Der Wert 0 reflektierte die subjektive Annahme, dass sich das entsprechende Merkmal nicht infolge des Lügens verändern würde. Weitere sieben Fragen dienten dazu, den Erfolg der experimentellen Manipulation zu überprüfen. Zwei davon bezogen sich auf die Gelegenheit zur Vorbereitung. So sollten die Versuchspersonen angeben, wie viele Minuten die lügende Person für die Vorbereitung ihrer Aussage aufgewendet hat und ob sie gut vorbereitet war. Die fünf Kontrollitems zur Wahrnehmung der Situation bezogen sich auf die Verwerflichkeit der Lüge, die Angst der lügenden Person überführt zu werden, deren Motivation erfolgreich zu überzeugen, den zu erwartenden negativen Konsequenzen bei einer Aufdeckung des Täuschungsversuches sowie auf die Wahrscheinlichkeit überführt zu werden. Ausgenommen der vermeintlichen Vorbereitungsdauer, die eine freie Zeiteinschätzung erforderte, wurden auch die Kontrollitems auf 7-stufigen Skalen von –3 bis +3 beurteilt.

Ergebnisse

Manipulation Checks

Der Erfolg der experimentellen Manipulationen wurde zunächst anhand von 3 x 2 ANOVAs mit den beiden Faktoren Relevanz der Täuschungssituation (gering/mittel/hoch) und Vorbereitungszeit (kurz/lang) überprüft. Als abhängige Variablen dienten die Kontrollitems. Bei signifikanten Effekten mehrstufiger Faktoren wurden zusätzlich paarweise, Bonferroni-adjustierte Mittelwertsvergleiche durchgeführt (alpha = .05 / Anzahl der Vergleiche). Die beiden Kontrollitems zur Vorbereitungszeit wurden wegen ihrer unterschiedlichen Antwortformate getrennt analysiert. Hingegen wurden die fünf Items zur Wahrnehmung der Situation aufgrund der gegebenen Interkorrelationen zusammengefasst, $.16 \leq r \leq .52$, alle p's $< .01$, Cronbach's alpha = .73.

Die vermeintliche Vorbereitungszeit wurde erfolgreich variiert. Probanden in den Bedingungen mit langer Vorbereitungszeit nahmen an, dass Lügner signifikant besser vorbereitet seien (M = 1.41) als Probanden in den Bedingungen mit kurzer Vorbereitungszeit (M = -0.97), F(1,234) = 126.96, p < .001, r = .59. Die Täuschungssituation hatte keinen signifikanten Haupteffekt auf die Beurteilung der Vorbereitungsgüte, F(2,234) = 1.75, p = .177, partielles eta^2 = .02. Jedoch war die Wechselwirkung zwischen den beiden Faktoren signifikant, F(2,234) = 4.08, p = .018, partielles eta^2 = .03. Es ergab sich ein einfacher Haupteffekt der Situation innerhalb der Bedingungen mit kurzer Vorbereitungszeit, F(2,234) = 4.62, p = .011, partielles eta^2 = .04, jedoch nicht innerhalb der Bedingungen mit langer Vorbereitungzeit, F(2,234) = 1.21, p = .299, partielles eta^2 = .01. In den Bedingungen kurzer Vorbereitungszeit nahmen die Probanden an, dass die lügende Person signifikant schlechter auf eine Falschaussage wegen eines Totschlagsdeliktes (M = -1.43) als wegen einer Verspätung (M = -0.35) vorbereitet sei. Die Beurteilung der Vorbereitungsgüte für die Affärensituation fiel dazwischen (M = -1.13), ohne sich signifikant von den beiden anderen Situationen zu unterscheiden. In den Bedingungen langer Vorbereitungzeit ergaben sich keine Unterschiede zwischen der Verspätungs- (M = 1.18), Affären- (M = 1.73) und Totschlagssituation (M = 1.33).

Auch in der Dauer der Vorbereitung zeigten sich erwartungsgemäße Unterschiede. Allerdings wurden bei den freien Zeiteinschätzungen einige Extremwerte sowie eine extreme Schiefe und Kurtosis der Verteilungen in verschiedenen Untersuchungsbedingungen festgestellt. Daher wurden die Werte vor der varianzanalytischen Überprüfung logarithmisch transformiert. Die im Folgenden zitierten Mittelwerte beziehen sich auf die Originalwerte. Da diese jedoch stark durch die vorliegenden Extremwerte beeinflusst werden, wird jeweils zusätzlich der Median aufgeführt. In Bedingungen mit langer Vorbereitung wurde die Dauer der Vorbereitung auf M = 50.34 Minuten (Mdn = 20.00), in Bedingungen mit kurzer Vorbereitung auf M = 16.52 Minuten (Mdn = 3.00) geschätzt, F(1,234) =

70.93, p < .001, r = .48. Zudem ergab sich ein signifikanter Effekt der Täuschungssituation auf die geschätzte Dauer der Vorbereitung, F(2,234) = 16.46, p < .001, partielles eta^2 = .12. Post-hoc Analysen zeigten, dass die Dauer der Vorbereitung in der Verspätungssituation signifikant geringer eingeschätzt wurde (M = 8.31, Mdn = 5.00) als in den beiden anderen Täuschungssituationen (M = 51.90, Mdn = 10.00 für das Affären-, M = 40.09, Mdn = 20.00 für das Totschlagsszenario). Die Wechselwirkung zwischen der Vorbereitungszeit und der Täuschungssituation war ebenfalls signifikant, F(2,234) = 11.85, p < .001, partielles eta^2 = .09. Unter den Bedingungen langer Vorbereitung wurde angenommen, dass die lügende Person sich signifikant kürzer auf ihre Falschaussage wegen einer Verspätung vorbereiten würde (M = 10.05, Mdn = 10.00) als wegen einer Affäre (M = 73.07, Mdn = 30.00) und wegen eines Totschlagsdeliktes (M = 67.90, Mdn = 40.00), F(2,234) = 27.48, p < .001, partielles eta^2 = .19. Es wurden jedoch keine signifikanten Unterschiede in der vermeintlichen Dauer der Vorbereitung zwischen der Affären- und Totschlagssituation festgestellt. Unter den Bedingungen kurzer Vorbereitungszeit zeigte sich hingegen kein einfacher Haupteffekt der Situation, F(2,234) = 0.83, p = .436, partielles eta^2 = .01. Es gab keine signifikanten Unterschiede in der geschätzten Vorbereitungsdauer für die Verspätungs- (M = 6.58, Mdn = 3.00), Affären- (M = 30.73, Mdn = 2.00) und Totschlagssituation (M = 12.28, Mdn = 4.00).

Die Kontrollitems zur Wahrnehmung der Situation wurden ursprünglich auf Skalen von –3 bis +3 beurteilt. Zur besseren Verständlichkeit wurde jedoch eine Konstante von vier hinzuaddiert, so dass positive Werte von 1 bis 7 resultierten. Danach wurden die Beurteilungen der Kontrollitems zur Wahrnehmung der Situation gemittelt. Infolgedessen indizieren hohe Werte eine negative Einschätzung der Situation. Hingegen verweisen geringe Werte darauf, dass die Situation als vergleichsweise harmlos aufgefasst wurde. Für die zu einer Variablen zusammengefassten Kontrollitems zur Wahrnehmung der Situation war erwartungsgemäß ein signifikanter Haupteffekt der Täuschungssituation

nachweisbar, $F(2,234) = 110.93$, $p < .001$, partielles $eta^2 = .49$. Post-hoc Analysen zeigten, dass sich die Mittelwerte aller drei Täuschungssituationen signifikant voneinander unterschieden. Wegen einer Verspätung zu lügen wurde als weniger negativ bewertet ($M = 4.42$) als wegen einer Affäre ($M = 5.81$) oder wegen eines Totschlagsdeliktes ($M = 6.26$). Es zeigte sich weder ein Haupteffekt der Vorbereitungszeit, $F(1,234) = 0.27$, $p = .60$, $r = .03$, noch eine Wechselwirkung der beiden variierten Faktoren auf die Wahrnehmung der Situation, $F(2,234) = 1.53$, $p = .22$, partielles $eta^2 = .01$.

Zusätzlich wurde der Erfolg der experimentellen Manipulation der Täuschungssituation unter Einschluss der Bedingung zum Sexualdelikt überprüft. Dazu wurde eine einfaktorielle ANOVA mit dem vierstufigen Between-Subjects-Faktor der Täuschungssituation gerechnet. Dabei wurden nur die Experimentalbedingungen berücksichtigt, in denen eine lange Vorbereitungszeit operationalisiert worden war. Die abhängige Variable war wiederum die Wahrnehmung der Situation, die sich aus der Zusammenfassung der entsprechenden fünf Kontrollitems ergab. Erneut war ein signifikanter Effekt der Täuschungssituation nachweisbar, $F(3,156) = 33.95$, $p < .001$, partielles $eta^2 = .40$. Die Verspätungssituation wurde weniger negativ beurteilt ($M = 4.56$) als die übrigen Szenarien. Zudem unterschied sich die Einschätzung der beiden rechtlich relevanten Szenarien zum Totschlags- ($M = 6.18$) und Sexualdelikt ($M = 5.38$) signifikant. Die Wahrnehmung der Affärensituation ($M = 5.84$) war mit beiden rechtlich relevanten Szenarien vergleichbar.

Vergleich subjektiver und objektiver Indikatoren

Die einzelnen Items wurden gemäß der Struktur der ARJS zu 13 Skalen zusammengefasst. Um die subjektiven Annahmen mit der in der Literatur berichteten objektiven Differenzierungskraft der Merkmale vergleichen zu können, wurden die Mittelwerte durch ihre jeweilige Standardabweichung dividiert. Die

resultierenden mittleren Effektstärken d der subjektiven Annahmen für jede Bedingung sind Tabelle 2.3 zu entnehmen.

Zudem wurden die Beurteilungen umgepolt, so dass positive Werte die subjektive Annahme repräsentieren, das Merkmal sei häufiger in wahren als in erfundenen Aussagen vorzufinden, was einer Interpretation als Wahrheitsindikator entspricht. Negative Werte verweisen darauf, dass ein Merkmal als Lügenindikator aufgefasst wird. Hingegen indiziert der Wert 0 die subjektive Annahme, dass das entsprechende Merkmal gleichermaßen bei wahren und erfundenen Aussagen vorzufinden sei und folglich keinerlei Differenzierungskraft aufweisen würde.

Um zu prüfen, ob sich die Effektstärken der subjektiven Annahmen signifikant von der Nullhypothese $d = 0$ unterscheiden, wurden one-sample t-Tests durchgeführt. Diese Analysen wurden gegen ein Bonferroni-korrigiertes Alpha-Niveau von $p < .0005$ abgesichert. Merkmale, die überzufällig häufig als Wahrheits- bzw. Lügenindikator wahrgenommen wurden, sind in Tabelle 2.3 fett gedruckt. Theoretisch sind die ARJS als Wahrheitsmerkmale aufzufassen. Positive Effektstärken der subjektiven Annahmen entsprechen daher der postulierten objektiven Differenzierungskraft der Merkmale, während negative Effektstärken Diskrepanzen aufzeigen.

Um die subjektiven Annahmen mit der empirisch ermittelten objektiven Differenzierungskraft der Merkmale zu vergleichen, wurden die Befunde von fünf Studien zur Validität der ARJS zusammengefasst (Barnier et al., 2005; Sporer, 1998; Sporer & Burghardt, 2004; Sporer et al., 2000; Sporer & Walther, 2006). Dazu wurden die in den einzelnen Studien berichteten Effektstärkemaße r einer Fisher's Z-Transformation unterzogen, danach ungewichtet gemittelt, in r-Werte zurücktransformiert und diese schließlich in d-Werte umgerechnet, $d = \sqrt{((4+r^2)/(1-r^2))}$. Die aus den Validierungsstudien abgeleiteten Effektstärkemaße der objektiven Täuschungsindikatoren sind ebenfalls in Tabelle 2.3 integriert.

Tabelle 2.3
Effektstärken d der subjektiven Täuschungsindikatoren für die sieben Experimentalbedingungen und Vergleich zur Kontrollgruppe sowie zu objektiven Indikatoren

	Verspätung		Affäre		Totschlag		Sexual-delikt	KG	d_{Obj}
	Vorbereitung		Vorbereitung		Vorbereitung		Vorbe-reitung		
ARJS-Skalen	Kurz	Lang	Kurz	Lang	Kurz	Lang	Lang		
Realismus/ Logische Struktur	0.27	-0.04a	0.30	**0.41**	**0.49**	0.14	0.24	**0.60**	0.07
Klarheit/ Lebendigkeit	**0.49**	**0.42**	0.18	**0.76**	**0.35**	0.26	**1.06a**	**0.33**	0.29
Details	0.05	0.15	0.01	-0.02	0.14	-0.08	**0.62**	0.26	0.51
Räumliche Details	**0.31**	**0.61**	0.23	0.30	0.20	-0.14a	**0.51**	**0.86**	0.03
Zeitliche Details	0.06	0.34	0.15	0.07	0.20	0.00	0.05	**0.37**	0.39
Sinneseindrücke	**0.51**	**0.66**	**0.76**	**0.64**	**1.40**	**1.05**	**1.95a**	**0.61**	0.01
Emotionen und Gefühle	-0.09	0.01	0.17	**0.38**	-0.06	-0.09	0.20	0.28	0.43
Gedanken	**-0.53**	**-0.41**	0.05	0.06	-0.29	**-0.59**	-0.13	0.02	0.34
Memorieren/ Gedächtnis	0.14	-0.22	-0.23	**-0.45**	**-0.33**	**-0.57**	0.08	-0.18	0.57
Nonverbale und verbale Interaktionen	0.25	**0.38**	0.21	**0.38**	0.23	0.04	**0.65**	0.25	0.30
Komplikationen/ ungewöhnliche Details	**-0.55**	**-0.51**	**-0.42**	-0.19	**-0.43**	-0.06	-0.32	-0.02	0.51
Fehler/ Sozial Unerwünschtes	-0.24	0.09	-0.08	0.03	-0.03	0.17	**0.46**	0.26	0.37
Persönliche Signifikanz	**-0.75**	**-0.42**	-0.19	**-0.44**	**-0.44**	**-0.94a**	-0.01	-0.27	0.31

Anm. Positive Effektstärken indizieren Interpretation als Wahrheitsindikator. Effektstärken, die sich signifikant von d = 0 unterscheiden sind fett gedruckt; Subskript a indiziert signifikante Unterschiede zur Kontrollgruppe; d_{Obj} sind die ungewichteten Mittelwerte von Barnier et al. (2005); Sporer, 1998; Sporer & Burghardt, 2004; Sporer et al. (2002); Sporer & Walther (2006).

Objektiv ergaben sich die stärksten Effekte für die Skalen Memorieren und Gedächtnis, Details sowie Komplikationen und ungewöhnliche Details. Die entsprechenden Merkmale waren bei wahren Aussagen stärker ausgeprägt als bei erfundenen. Im Gegensatz dazu ergaben sich für die subjektiven Annahmen hinsichtlich der Skalen Memorieren und Gedächtnis sowie Komplikationen und ungewöhnliche Details meist negative Effektstärken. Die Merkmale wurden demnach eher mit erfundenen Aussagen assoziiert, wobei sich dies nicht gegen die Nullhypothese absichern ließ. Für die Skala Details zeigten sich oft positive Effektstärken. Allerdings ließ sich die subjektive Annahme, dass die Skala mit wahren Aussagen assoziiert sei, nur in der Bedingung zum Sexualdelikt gegen die Nullhypothese absichern. Die Validität der Skalen Sinneseindrücke, räumliche Details sowie Realismus und logische Struktur wurde bislang nicht erfolgreich nachgewiesen. Die Effektstärken für die subjektiven Annahmen fielen jedoch recht hoch aus. Meist wurden positive Effektstärken ermittelt, d.h. die Merkmale wurden eher mit wahren Aussagen assoziiert. Für die Skala Fehler und sozial Unerwünschtes zeigen die vorliegenden Referenzstudien, dass die entsprechenden Merkmale objektiv häufiger bei wahren Aussagen als bei erfundenen Aussagen vorzufinden sind. Subjektiv wurde jedoch in mehreren Bedingungen angenommen, dass sich die entsprechenden Merkmale nicht infolge des Lügens verändern würden. In anderen Bedingungen fielen die ermittelten Effektstärken positiv aus, d.h. die Merkmale wurden dem objektiven Forschungsstand entsprechend eher mit wahren Aussagen assoziiert. Lediglich in der Verspätungssituation mit kurzer Vorbereitungszeit zeigte sich eine geringe negative Effektstärke hinsichtlich der Skala Fehler und sozial Unerwünschtes, die sich jedoch nicht gegen die Nullhypothese absichern ließ.

Effekte der experimentellen Variation auf die ARJS

Anhand einer 3 x 2 MANOVA wurde der Einfluss der experimentellen Variation der Relevanz der Täuschungssituation (gering/mittel/hoch) und der Vorbereitungszeit (kurz/lang) auf die subjektiven Annahmen hinsichtlich der 13 ARJS-Skalen überprüft. Die Kontrollgruppe sowie die Experimentalgruppe, die ihre Angaben auf das Szenario eines Sexualdeliktes bezog, wurden zunächst von der Analyse ausgeschlossen.

Es zeigte sich ein multivariat signifikanter Haupteffekt der Täuschungssituation, multivariates $F(26,444) = 2.61$, Wilks' Lambda = .75, $p < .001$, partielles $eta^2 = .13$. Tabelle 2.4 ist zu entnehmen, dass dieser Effekt univariat auf situationsspezifische Beurteilungen der drei Skalen räumliche Details, Gedanken sowie Memorieren und Gedächtnis zurückzuführen ist, alle Fs$(2,234) \geq 3.28$, ps $< .05$, partielle eta^2s $\geq .03$. Für die Skalen räumliche Details sowie Memorieren und Gedächtnis verwiesen post-hoc Analysen auf signifikante Unterschiede zwischen dem Verspätungs- und Totschlagsszenario. Für die Skala Gedanken hingegen waren die subjektiven Annahmen des Verspätungs- und Totschlagsszenarios vergleichbar, während sie sich signifikant von der Beurteilung in der Affärensituation unterschieden.

Ein Haupteffekt der Vorbereitungszeit war nicht nachweisbar, multivariates $F(13,222) = 1.17$, Wilks' Lambda = .94, $p = .306$, partielles $eta^2 = .06$. Auch zeigte sich keine Wechselwirkung zwischen der Vorbereitungszeit und der Täuschungssituation, multivariates $F(26,444) = 1.13$, Wilks' Lambda = .88, $p = .301$, partielles $eta^2 = .06$. Dies galt wider Erwarten auch auf univariater Ebene für die Skala Details, $F(2,234) = 0.65$, $p = .522$, partielles $eta^2 = .01$.

Tabelle 2.4

Effektstärken d der subjektiven Indikatoren für die Verspätungs-, Affären- und Totschlagsituation sowie F- und p-Werte der univariaten Analysen

ARJS-Skalen	Verspätung	Affäre	Totschlag	F(2,234)	p
Realismus/ Logische Struktur	0.10	0.35	0.31	1.21	.300
Klarheit/ Lebendigkeit	0.46	0.45	0.31	0.71	.494
Details	0.11	-0.01	0.03	0.25	.781
Räumliche Details	0.47a	0.26ab	0.03b	**3.28**	**.039**
Zeitliche Details	0.14	0.08	0.09	0.14	.868
Sinneseindrücke	0.59	0.70	1.21	2.13	.122
Emotionen und Gefühle	-0.04	0.27	-0.08	2.88	.058
Gedanken	-0.47a	0.05b	-0.44a	**6.63**	**.002**
Memorieren/ Gedächtnis	-0.04a	-0.33ab	-0.46b	**3.71**	**.026**
Nonverbale und verbale Interaktionen	0.32	0.29	0.12	0.80	.451
Komplikationen/ ungewöhnliche Details	-0.53	-0.29	-0.22	1.83	.163
Fehler/Sozial Unerwünschtes	-0.08	-0.03	0.06	0.42	.656
Persönliche Signifikanz	-0.56	-0.31	-0.66	1.60	.204

Anm. N = 80 pro Situation. Werte mit unterschiedlichen Subskripta unterscheiden sich signifikant.

Um auch das Szenario zum Sexualdelikt bei den Analysen zu berücksichtigen, wurde eine weitere MANOVA gerechnet. Als vierstufiger Faktor diente die Täuschungssituation (Verspätung, Affäre, Totschlag, Sexualdelikt), wobei nur die Bedingungen betrachtet wurden, in denen ein hohes Ausmaß an Vorbereitungszeit realisiert wurde. Diese einfaktorielle MANOVA erwies sich ebenfalls als signifikant, multivariates $F(39,427) = 2.55$, Wilks' Lambda = .55, $p < .001$, partielles $eta^2 = .19$. Ging die Bedingung zum Sexualdelikt in die Analysen mit ein, wurden wesentlich mehr Skalen situationsspezifisch beurteilt (Tabelle 2.5). Unterschiede zwischen den vier Täuschungssituationen ließen sich seltener auf die Richtung als auf die Stärke der angenommenen Effekte zurückführen.

Post-hoc Analysen verwiesen insbesondere auf Unterschiede in den subjektiven Annahmen für die beiden rechtlich relevanten Szenarien zum Totschlags- und Sexualdelikt. Für die Skalen Klarheit und Lebendigkeit, Details, räumliche Details und Sinneseindrücke ergaben sich stärkere Effekte für das Szenario zum Sexualdelikt als zum Totschlagsdelikt. Umgekehrt resultierten hinsichtlich der Skalen Memorieren und Gedächtnis sowie persönliche Signifikanz stärkere Effekte für das Totschlagsszenario. Der univariat signifikante Effekt der Täuschungssituation hinsichtlich der Skala Komplikationen und ungewöhnliche Details ließ sich auf unterschiedliche subjektive Annahmen in der Verspätungssituation und dem Szenario zum Sexualdelikt zurückführen. Für die Skala Gedanken zeigten sich trotz eines univariaten Effekts der Täuschungssituation post-hoc keine signifikanten Unterschiede zwischen den einzelnen Untersuchungsbedingungen.

Tabelle 2.5

Effektstärken d der subjektiven Indikatoren für die Experimentalbedingungen mit langer Vorbereitungszeit sowie F- und p-Werte der univariaten Analysen

ARJS-Skalen	Verspätung	Affäre	Totschlag	Sexualdelikt	F(3,156)	p
Realismus/ Logische Struktur	-0.04	0.41	0.14	0.24	1.14	.334
Klarheit/ Lebendigkeit	0.42ab	0.76ab	0.26a	1.06b	**4.11**	**.008**
Details	0.15ab	-0.02a	-0.08a	0.62b	**3.58**	**.015**
Räumliche Details	0.61a	0.30ab	-0.14b	0.51a	**4.42**	**.005**
Zeitliche Details	0.34	0.07	-0.00	0.05	0.55	.652
Sinneseindrücke	0.66a	0.64a	1.05a	1.95b	**6.18**	**.001**
Emotionen und Gefühle	0.01	0.38	-0.09	0.20	1.57	.199
Gedanken	-0.41	0.06	-0.59	-0.13	**2.83**	**.040**
Memorieren/ Gedächtnis	-0.22ab	-0.45ab	-0.57a	0.08b	**3.39**	**.020**
Nonverbale und verbale Interaktionen	0.38	0.38	0.04	0.65	2.14	.097
Komplikationen/ ungewöhnliche Details	-0.51a	-0.19ab	-0.06ab	0.32b	**4.16**	**.007**
Fehler / Sozial Unerwünschtes	0.09	0.03	0.17	0.46	1.45	.230
Persönliche Signifikanz	-0.42ab	-0.44ab	-0.94a	-0.01b	**3.83**	**.011**

Anm. N = 40 pro Situation, Werte mit unterschiedlichen Subskripta unterscheiden sich signifikant.

Vergleich der Experimentalgruppen mit der Kontrollgruppe

Schließlich wurde anhand von Dunnett's t-Tests überprüft, ob sich die subjektiven Annahmen in den einzelnen Untersuchungsbedingungen von denen der Kontrollgruppe unterschieden (Tabelle 2.3).

Lediglich für vier Skalen waren signifikante Unterschiede nachweisbar. Hinsichtlich der Skalen Klarheit und Lebendigkeit sowie Sinneseindrücke ergaben sich für das Szenario zum Sexualdelikt stärkere Effekte als für die Kontrollbedingung. Der Skala persönliche Signifikanz wurde in der Totschlagssituation mit langer Vorbereitungszeit eine stärkere, der Skala räumliche Details hingegen eine geringere Differenzierungskraft zugesprochen als in der Kontrollbedingung. Schließlich zeigten sich unterschiedliche subjektive Annahmen für die Skala Realismus und logische Struktur zwischen der Alltagssituation mit langer Vorbereitungszeit und der Kontrollbedingung.

Diskussion

Im Rahmen der vorliegenden Fragebogenstudie wurden subjektive Annahmen zu Täuschungsindikatoren für inhaltliche Aussagemerkmale erfasst. Dabei wurden die Gelegenheit zur Vorbereitung sowie die spezifische Täuschungssituation über die Vorgabe fiktiver Szenarien variiert.

Wahrgenommene Unterschiede zwischen den Szenarien

Das Ausmaß an Vorbereitungszeit wurde erfolgreich variiert. Ursprünglich sollten die Szenarien in den Bedingungen mit geringer Vorbereitungszeit nahelegen, dass keinerlei Gelegenheit zur Vorbereitung bestand. Die Ergebnisse des Manipulation Checks ergaben jedoch, dass der Median der vermeintlichen Vorbereitungsdauer bei 3 Minuten lag. Die resultierenden subjektiven Annahmen lassen sich demnach nicht auf gänzlich unvorbereitete Aussagen generalisieren. Dennoch waren die Unterschiede zwischen Bedingungen mit kurzer und langer Vorbereitungszeit sehr deutlich. So lag der Median der vermeintlichen Vorbereitungsdauer für die Bedingungen mit langer Vorbereitung bei 20 Minuten.

Auch die Täuschungssituationen wurden intentionsgemäß wahrgenommen. Es wurde von der Verspätungs- über die Affärensituation hin zum Totschlagsdelikt als zunehmend negativer beurteilt zu lügen. Hinsichtlich der beiden rechtlich relevanten Szenarien unter der Bedingung langer Vorbereitungszeit wurde es als deutlich negativer erachtet wegen eines Totschlagsdeliktes zu lügen als wegen eines Sexualdeliktes. Dies ist auf eine vergleichsweise positive Bewertung des Szenarios zum Sexualdelikt zurückzuführen. Aufgrund spezifischer Eigenarten des verwendeten Szenarios erscheint diese Wahrnehmung nachvollziehbar. So wurde expliziert, dass tatsächlich ein Sexualdelikt stattgefunden hat, sich jedoch die falsche Person als Opfer ausgibt, um das tatsächliche Opfer zu schützen und den Täter zu überführen. Dieses Aussagemotiv hat möglicherweise zu der verhältnismäßig positiven Einschätzung der Situation beigetragen. Wenn nur die Bedingungen betrachtet wurden, in denen eine lange Vorbereitungszeit realisiert wurde, war die Wahrnehmung der Affärensituation wiederum mit beiden rechtlich relevanten Szenarien vergleichbar. Möglicherweise wurde die Affärensituation recht negativ beurteilt, weil sich die studentische Stichprobe eher vorstellen konnte, eine Situation selbst zu erleben, die dem Affärenszenario ähnelt, als den Straftatszenarien.

Diskussion der Untersuchungshypothesen

Die meisten Aussagemerkmale wurden gemäß ihrer objektiven Differenzierungskraft auch subjektiv als Wahrheitsindikatoren gewertet. Den bisherigen Forschungsstand zu subjektiven Indikatoren replizierend, zeigte sich, dass wahre Aussagen mit einem höheren Ausmaß an Realismus und logischer Struktur, mehr Verweisen auf nonverbale und verbale Interaktionen und Details assoziiert wurden. Räumlichen Details wurde eine stärkere Differenzierungskraft zugesprochen als zeitlichen, obwohl nach den bisherigen Befunden zu den ARJS objektiv letztere valider zu sein scheinen. Die Schilderung von Gedanken wurde

hingegen eher mit erfundenen Aussagen assoziiert. Dies entspricht der Vorhersage des RÜ-Ansatzes zur objektiven Differenzierungskraft, die jedoch durch die vorliegenden Referenzstudien nicht gestützt wurde. Insgesamt ist der bisherige Forschungsstand zu diesem Aussagemerkmal als widersprüchlich zu bewerten (vgl. Masip et al., 2005; Sporer, 2004), was auf unterschiedliche Operationalisierungen des Merkmals kognitive Operationen zurückzuführen sein dürfte.

Hinsichtlich von Fehlern und sozial Unerwünschtem wird angenommen, dass diese Merkmale häufiger in wahren Aussagen vorzufinden sind, weil sie bei Falschaussagen gezielt vermieden werden (DePaulo et al., 2003; Sporer, 1996/1998/2004; Steller & Köhnken, 1989). Diese Argumentation erfordert den Nachweis, dass die Merkmale stereotypen Vorstellungen von Täuschung entsprechen. Bisherige Studien zur Erfassung subjektiver Annahmen von Täuschungskorrelaten erzielten jedoch unterschiedliche Ergebnisse. Auch in der vorliegenden Untersuchung ergaben sich je nach Bedingung positive oder negative Effektstärken für die Skala Fehler und sozial Unerwünschtes der ARJS. Lediglich für die Verspätungssituation, in der keine Gelegenheit zur Vorbereitung bestand, wurde diese Skala tendenziell als Lügenindikator gewertet. Folglich unterstützen die vorliegenden Untersuchungsbefunde nicht die Argumentation, dass die Validität dieser Merkmale auf stereotype Täuschungsannahmen zurückzuführen sei (Hypothese 1a). Allerdings ist darauf hinzuweisen, dass die ursprünglichen CBCA-Merkmale für die ARJS aufgrund gedächtnistheoretischer und sozialpsychologischer Überlegungen modifiziert und präzisiert wurden. Beispielsweise wurden die CBCA-Merkmale „Eingeständnis von Erinnerungslücken“ und „Selbstbelastungen“ zu den ARJS-Merkmalen „Erinnerungslücken hinsichtlich einiger Details zugeben“ und „unschmeichelhafte Handlungen aufgrund persönlicher Schwächen, Irrtümer und Fehler zugeben“ umformuliert. Die CBCA- und ARJS-Merkmale sind daher nicht direkt vergleichbar.

Die CBCA-Merkmale Schilderungen von Komplikationen im Handlungsverlauf und von ausgefallenen Einzelheiten wurden in bisherigen Fragebogenuntersuchungen oftmals mit erfundenen Aussagen assoziiert. In Übereinstimmung verweisen auch die im Rahmen der vorliegenden Untersuchung ermittelten negativen Effektstärken für die ARJS-Skala Komplikationen und ungewöhnliche Details darauf, dass die dazugehörigen Merkmale als Lügenindikatoren aufgefasst werden (Hypothese 1b).

Insgesamt waren die Diskrepanzen zwischen subjektiven und objektiven Indikatoren für inhaltliche Aussagemerkmale geringer, als es für non- und paraverbale Merkmale berichtet wurde (Breuer et al., 2005). Die realistischere Einschätzung inhaltlicher Aussagemerkmale ist möglicherweise darauf zurückzuführen, dass Personen ihr verbales Verhalten bewusster wahrnehmen und stärker reflektieren als ihr para- und nonverbales Verhalten. Dennoch wurde für inhaltliche Aussagemerkmale eine stärkere objektive Differenzierungskraft nachgewiesen als für para- und nonverbale Merkmale (z.B. DePaulo et al., 2003). Obwohl für inhaltliche Aussagemerkmale subjektive und objektive Indikatoren eher korrespondieren als für para- und nonverbale Merkmale, beeinträchtigt dies offenbar nicht die Validität der inhaltlichen Aussagemerkmale. Anscheinend gelingt es Personen nicht die Inhalte ihrer Aussagen gemäß ihrer subjektiven Annahmen auszurichten.

Einige Autoren führten aus, dass Personen bei der Beurteilung von Täuschungsindikatoren möglicherweise an Situationen mit hoher Relevanz denken (Hypothese 5). Bei einem Vergleich der subjektiven Annahmen für die einzelnen Experimentalgruppen mit denen der Kontrollgruppe zeigten sich keine bzw. kaum Unterschiede zu den Affären- und Verspätungsszenarien. Die meisten Unterschiede zeigten sich zwischen der Kontrollgruppe und den Bedingungen zum Totschlags- oder Sexualdelikt. Beispielsweise wurde die Differenzierungskraft der Skalen Klarheit und Lebendigkeit sowie Sinneseindrücke für das Sexualdelikt höher eingeschätzt als bei fehlender Situationsvorgabe.

Hinsichtlich der persönlichen Signifikanz ergab sich ein stärkerer Effekt für das Totschlagsszenario mit langer Vorbereitungszeit als für die Kontrollgruppe. Folglich scheint die Kontrollgruppe zumindest nicht an Situationen gedacht zu haben, die mit einem der beiden verwendeten Straftatsszenarien vergleichbar wären.

Lakhani und Taylor (2003) verwendeten in ihrer Untersuchung ebenfalls fiktive Szenarien, um subjektive Annahmen über Täuschungskorrelate zu untersuchen. Dabei wurden die Bedingungen einer geringfügig bzw. hoch relevanten Täuschungssituation über jeweils zwei verschiedene Szenarien operationalisiert. Als Szenarien hoher Relevanz wurden ein selbst-verschuldetes Verkehrsdelikt mit Todesfolge sowie fahrlässiges Verhalten, das Einbruch, Diebstahl und Vandalismus zur Folge hatte, dargestellt. Die Autoren berichteten, dass sich die subjektiven Annahmen für die Szenarien mit vergleichbarer Relevanz nicht voneinander unterschieden, d.h. die Ergebnisse waren über die konkreten Täuschungssituationen hinaus vergleichbar.

Im Gegensatz dazu unterschieden sich in der vorliegenden Untersuchung die subjektiven Annahmen für die beiden rechtlich relevanten Szenarien mit langer Vorbereitungszeit in sechs Skalen. Dies lässt im Einklang mit Niehaus et al. (2005) vermuten, dass Sexualdelikte besondere Eigenarten aufweisen, die in spezifischen subjektiven Annahmen resultieren (Hypothese 6). Das Szenario zum Sexualdelikt war das einzige, bei dem vorgegeben wurde, die lügende Person sei über das tatsächliche Ereignis informiert und brauche es lediglich nachzuerzählen. Aus theoretischen Überlegungen sind solche Falschaussagen besonders problematisch, da sie keine freien Erfindungen erfordern. Daher ist anzunehmen, dass für solche Situationen objektive Unterschiede zwischen wahren und falschen Aussagen reduziert sind. Beispielsweise fanden Sporer et al. (2000) bei wahren Aussagen mehr Verweise auf zeitliche Informationen als bei frei erfundenen. Basierten die Falschaussagen jedoch auf einer modifizierten Darstellung persönlicher Erlebnisse, ließen sich diesbezüglich keine

Unterschiede mehr zu wahren Aussagen nachweisen. Vor diesem Hintergrund überrascht es, dass die subjektiven Annahmen in der vorliegenden Studie für die Bedingung zum Sexualdelikt vergleichsweise stark ausfielen.

Hypothesenkonform war ein Haupteffekt der Situation nachweisbar (Hypothese 2). Betrachtet man die vier Bedingungen, in denen eine lange Vorbereitungszeit realisiert wurde, so wurden acht Skalen situationsspezifisch beurteilt. Die subjektiven Annahmen für das Affären- und Totschlagsszenario waren dabei stets vergleichbar. Vor dem Hintergrund, dass diese beiden Szenarien hinsichtlich der meisten Kontrollitems ähnlich wahrgenommen wurden, ist dies nicht weiter verwunderlich. Insbesondere für das Sexualdelikt zeigten sich situationsspezifische subjektive Annahmen. Beispielsweise wurden Verweise auf Komplikationen für das Verspätungsszenario als Lügenindikator aufgefasst, für das Szenario zum Sexualdelikt hingegen als Wahrheitsindikator gewertet. Beim Verspätungsszenario wurde vorgegeben, dass ein beruflicher Anruf der Grund für die Verspätung sei. Vermutlich erachteten die Probanden es als typische Ausrede für Verspätungssituationen auf unvorhersehbare Ereignisse (wie eine Zugverspätung, Unfall auf der Autobahn) zu verweisen. Die ARJS-Skalen Memorieren und Gedächtnis sowie Verweise auf die persönliche Signifikanz des Ereignisses wurden eher mit erfundenen Aussagen assoziiert. Für das Szenario zum Sexualdelikt wurde jedoch angenommen, dass diese Skalen nicht zwischen wahren und erfundenen Aussagen differenzieren, obwohl sie objektiv zur Differenzierung geeignet zu sein scheinen.

Insgesamt scheinen Versuchspersonen durchaus situative Besonderheiten bei ihren subjektiven Annahmen über Täuschungsindikatoren zu berücksichtigen. Allerdings waren in der vorliegenden Untersuchung weder ein Haupteffekt der Vorbereitungszeit (Hypothese 1) noch die postulierte Wechselwirkung zwischen der Vorbereitungszeit und der Täuschungssituation nachweisbar (Hypothese 4a-c). Aufgrund der Analysen zum Manipulation Check sind Unzulänglichkeiten in der Operationalisierung als Ursache hierfür weitestgehend auszuschließen. Auch in

der Fragebogenuntersuchung von Taylor und Vrij (2000) zeigte sich kein Effekt der variierten kognitiven Komplexität der Lüge auf subjektive Annahmen zu Täuschungsindikatoren. Ein geringe kognitive Komplexität der Falschaussage wurde über das Fehlen, eine hohe über das Vorliegen gegenteiliger Zeugenaussagen operationalisiert. Integriert man die vorliegenden Untersuchungsbefunde und die von Taylor und Vrij (2000), so ließe sich schlussfolgern, dass Personen in ihren subjektiven Annahmen wenig sensitiv für unterschiedliche Schwierigkeitsgrade beim Lügen sind (sei es durch die Gelegenheit zur Vorbereitung oder aufgrund von externen Beweismitteln). Auch der Befund, dass für das Szenario zum Sexualdelikt die objektive Differenzierungskraft der meisten Merkmale überschätzt wurde, ließe sich darüber erklären. Die Versuchspersonen ignorierten möglicherweise, dass die aussagende Person lediglich die wahre Aussage des tatsächlichen Opfers nachzuerzählen brauchte und dass dies einfacher ist als eine Falschaussage frei zu erfinden. Allerdings gaben Personen in einer freien Befragung von Granhag und Strömwall (2000) an, bei der Beurteilung von Aussagen durchaus zu berücksichtigen, ob diese spontan vorgebracht werden, oder den Eindruck vermitteln, vorbereitet oder einstudiert worden zu sein.

Zusammenfassend scheinen Personen für einige Moderatoren von Täuschungsindikatoren sensitiv zu sein, für andere wiederum nicht. In der vorliegenden Untersuchung zeigte sich kein Effekt der variierten Vorbereitungszeit, während die inhaltlichen Merkmale in Abhängigkeit vom Gegenstand der Lüge durchaus unterschiedlich bewertet wurden. Welche Merkmale der Täuschungssituationen für Variationen in den subjektiven Annahmen verantwortlich sind, bleibt allerdings weitestgehend unklar. Zumindest war keine kontinuierliche Zu- oder Abnahme der Effektstärken in Abhängigkeit von der Relevanz der Lüge zu beobachten. Die meisten Effektstärken unterschieden sich lediglich in ihrer Größenordnung, nicht aber in ihrer Richtung. Nur vereinzelt wurden Indikatoren in einer spezifischen Situation mit wahren, in einer anderen

hingegen mit erfundenen Aussagen assoziiert. Dies galt beispielsweise für die Skala Fehler und sozial Unerwünschtes. Deren Validität lässt sich demnach nicht grundsätzlich darauf zurückführen, dass die Merkmale stereotypen Vorstellungen von Täuschung entsprechen und deswegen bei Falschaussagen gezielt vermieden werden. Allerdings sind weitere Untersuchungen erforderlich, um diesen Befund abzusichern. Insbesondere Paradigmen, die kein gebundenes Antwortformat nutzen, könnten die vorliegende Untersuchung sinnvoll ergänzen.

Literatur

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar? A meta-analysis of individual differences in detecting deception. Forensic Examiner, 15, 6-11.

Akehurst, L., Köhnken, G., Vrij, A., & Bull, R. (1996). Lay persons' and police officers' beliefs regarding deceptive behaviour. Applied Cognitive Psychology, 10, 461-471.

Alonso-Quecuty, M. L. (1992). Deception detection and reality monitoring: A new answer to an old question? In F. Lösel, D. Bender, & T. Bliesener (Eds.), Psychology and law: International perspectives (pp. 328-335). New York: Walter de Gruyter.

Anderson, D. E., DePaulo, B. M., Ansfield, M. E., Tickle, J. J., & Green, E. (1999). Beliefs about cues to deception: Mindless stereotypes or untapped wisdom? Journal of Nonverbal Behavior, 23, 67-89.

Barnier, A. J., Sharman, S. J., McKay, L., & Sporer, S. L. (2005). Discriminating adults' genuine, imagined, and deceptive accounts of positive and negative childhood events. Applied Cognitive Psychology, 19, 985-1001.

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. Personality and Social Psychology Review, 10, 214-234.

Breuer, M. M., Sporer S. L., & Reinhard, M. A. (2005). Subjektive Indikatoren von Täuschung: Die Bedeutung von Situation und Gelegenheit zur Vorbereitung. Zeitschrift für Sozialpsychologie, 36, 189-201.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. Psychological Bulletin, 129, 74-112.

Feeley, T. H., & deTurck, M. A. (1998). The behavioral correlates of sanctioned and unsanctioned deceptive communication. Journal of Nonverbal Behavior, 22, 189-204.

Feeley, T. H., & Young, M. J. (2000). Self-reported cues about deceptive and truthful communication: The effects of cognitive capacity and communicator veracity. Communication Quaterly, 48, 101-119.

Fiedler, K. (1989a). Lügendetektion aus alltagspsychologischer Sicht. Psychologische Rundschau, 40, 127-140.

Fiedler, K. (1989b). Suggestion and credibility: Lie detection based on content-related cues. In V. A. Gheorghin, P. Netter, H. J. Eysenck, & R. Rosenthal (Eds.), Suggestion and suggestiblity: Theory and research (pp. 323-335). Berlin: Springer.

Forrest, J. A., Feldman, R. S., & Tyler, J. M. (2004). When accurate beliefs lead to better lie detection. Journal of Applied Social Psychology, 34, 764-780.

Freedman, J. L., Adam, E. K., Davey, S. A., & Koegel, C. J. (1996). The impact of a statement: More detail does not always help. Legal and Criminological Psychology, 1, 117-130.

Global Deception Research Team (2006). A world of lies. Journal of Cross-Cultural Psychology, 37, 60-74.

Granhag, P. A., Andersson, L. O., Strömwall, L. A., & Hartwig, M. (2004). Imprisoned knowledge: Criminals' beliefs about deception. Legal and Criminological Psychology, 9, 103-119.

Granhag, P. A., & Strömwall, L. A. (2000). Effects of preconceptions on deception detection and new answers to why lie-catchers often fail. Psychology, Crime, and Law, 6, 197-218.

Granhag, P. A., & Strömwall, L. A. (2001). Deception detection: Interrogators' and observers' decoding of consecutive statements. The Journal of Psychology, 135, 603-620.

Granhag, P. A., Strömwall, L. A., & Hartwig, M. (2005). Granting asylum or not? Migration board personnel's beliefs about deception. Journal of Ethnic and Migration Studies, 31, 29-50.

Granhag, P. A., Strömwall, L. A., & Landström, S. (2006). Children recalling an event repeatedly: Effects on RM and CBCA scores. Legal and Criminological Psychology, 11, 81-98.

Greuel, L. (1992). Police officers' beliefs about cues associated with deception in rape cases. In F. Lösel, D. Bender, & T. Bliesener (Eds.), Psychology and Law (pp. 234-239). Berlin: deGruyter.

Hartwig, M., Granhag, P. A., Strömwall, L. A., & Andersson, L. O. (2004). Suspicious minds: Criminals' ability to detect deception. Psychology, Crime, and Law, 10, 83-95.

Hernandez-Fernaud, E., & Alonso-Quecuty, M. L. (1997). The cognitive interview and lie detection: A new magnifying glass for Sherlock Holmes? Applied Cognitive Psychology, 11, 55-68.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. Psychological Review, 88, 67-85.

Johnson, M. K., & Suengas, A. G. (1989). Reality monitoring judgments of other people's memories. Bulletin of the Psychonomic Society, 27, 107-110.

Köhnken, G. (1990). Glaubwürdigkeit. Untersuchungen zu einem psychologischen Konstrukt. München: Psychologie Verlags Union.

Kraut, R. E. (1978). Verbal and nonverbal cues in the perception of lying. Journal of Personality and Social Psychology, 36, 380-391.

Lakhani, M., & Taylor, R. (2003). Beliefs about the cues to deception in high- and low-stake situations. Psychology, Crime, and Law, 9, 357-368.

Landström, S., Granhag, P. A., & Hartwig, M. (2005). Witnesses appearing live versus on video: Effects on observers' perception, veracity assessments and memory. Applied Cognitive Psychology, 19, 913-933.

Mann, S., & Vrij, A. (2006). Police officers' judgements of veracity, tenseness, cognitive load and attempted behavioural control in real-life police interviews. Psychology, Crime, and Law, 12, 307-319.

Mann, S., Vrij, A., & Bull, R. (2002). Suspects, lies, and videotape: An analysis of authentic high-stake liars. Law and Human Behavior, 26, 365-376.

Mann, S., Vrij, A., & Bull, R. (2004). Detecting true lies: Police officers' ability to detect suspects' lies. Journal of Applied Psychology, 89, 137-149.

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. Psychology, Crime, and Law, 11, 99-122.

Niehaus, S., Krause, A., & Schmidke, J. (2005). Täuschungsstrategien bei der Schilderung von Sexualstraftaten. Zeitschrift für Sozialpsychologie, 36, 175-187.

Reinhard, M. A., Burghardt, K., Sporer, S. L., & Bursch, S. E. (2002). Alltagsvorstellungen über inhaltliche Kennzeichen von Lügen: Selbstberichtete Begründungen bei konkreten Glaubwürdigkeitsurteilen. Zeitschrift für Sozialpsychologie, 33, 169-180.

Ruby, C. L., & Brigham, J. C. (1998). Can criteria-based content analysis distinguish between true and false statements of African-American speakers? Law and Human Behavior, 22, 369-388.

Sporer, S. L. (1996/1998/2004). The Aberdeen Report Judgment Scales (ARJS). Definitions and answer sheets. Unpublished Questionnaire. University of Aberdeen, Scotland; University of Giessen, Germany.

Sporer, S. L. (1997a). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. Applied Cognitive Psychology, 11, 373-397.

Sporer, S. L. (1998, March). Detecting deception with the Aberdeen Report Judgment Scales (ARJS): Theoretical development reliability and validity. Paper presented at the Biennial Meeting of the American Psychology-Law Society, Redondo Beach, CA.

Sporer, S. L. (2004). Reality monitoring and detection of deception. In P.-A. Granhag & L. Strömwall (Eds.), Deception detection in forensic contexts (pp. 64-102). Cambridge: University Press.

Sporer, S. L., & Burghardt, K. (2004, March). Truth detection with the Aberdeen Report Judgment Scales: The role of planning and rehearsal. Paper presented at the Biennial Meeting of the American Psychology-Law Society, Phoenix, AZ.

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

Sporer, S. L., & Küpper, B. (2004). Fantasie und Wirklichkeit: Erinnerungsqualitäten von erlebten und erfundenen Geschichten. Zeitschrift für Psychologie, 212, 135-151.

Sporer, S. L., Samweber, M. C., & Stucke, T. S. (2000, March). Twisting the outcome: Discriminating distorted truths from factually experienced events. Paper presented at the Biennial Meeting of the American Psychology-Law Society, New Orleans, LA.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. Applied Cognitive Psychology, 20, 421-446.

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. Psychology, Public Policy, and Law, 13, 1-34.

Sporer, S. L., & Sharman, S. J. (2006). Should I believe this? Reality monitoring of accounts of self-experienced and invented recent and distant autobiographical events. Applied Cognitive Psychology, 20, 1-18.

Sporer, S. L., & Walther, A. (2006, March). Truth detection by content cues: General vs. specific questions. Paper presented at the Meeting of the American Psychology-Law Society in Petersburg, FL.

Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille (Ed.), Credibility assessment (pp. 135-154). Deventer: Kluwer Academic Publishers.

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

Steller, M., Wellershaus, P., & Wolf, T. (1992). Realkennzeichen in Kinderaussagen: Empirische Grundlage der kriterienorientierten Aussageanalyse. Zeitschrift für Experimentelle und Angewandte Psychologie, 39, 151-170.

Strömwall, L. A., & Granhag, P. A. (2003). How to detect deception? Arresting the beliefs of police officers, prosecutors and judges. Psychology, Crime, and Law, 9, 19-36.

Strömwall, L. A., Granhag, P. A., & Hartwig, M. (2004). Professionals' beliefs about deception. In P. A. Granhag & L. A. Strömwall (Eds.), The detection of deception in forensic contexts (pp. 229-250). Cambridge: Cambridge University Press.

Strömwall, L. A., Granhag, P. A., & Landström, S. (2007). Children's prepared and unprepared lies: Can adults see through their strategies? Applied Cognitive Psychology, 21, 457-471.

Suengas, A. G., & Johnson, M. K. (1988). Qualitative effects of rehearsal on memories for perceived and imagined complex events. Journal of Experimental Psychology: General, 117, 377-389.

Taylor, R., & Hick, R. F. (2007). Believed cues to deception: Judgments in self-generated trivial and serious attention. Legal and Criminological Psychology, 12, 321-331.

Taylor, R., & Hill-Davies, C. (2004). Parents' and non-parents' beliefs about the cues to deception in children. Psychology, Crime, and Law, 10, 455-464.

Taylor, R., & Vrij, A. (2000). The effects of varying stake and cognitive complexity on beliefs about the cues to deception. International Journal of Police Science and Management, 3, 111-123.

Vrij, A. (2000/2008). Detecting lies and deceit: The psychology of lying and the implications for professional practice. Chichester: John Wiley.

Vrij, A. (2005). Criteria-based content analysis: A qualitative review of the first 37 studies. Psychology, Public Policy, and Law, 11, 3-41.

Vrij, A., Akehurst, L., & Knight, S. (2006). Police officers', social workers' and the general public's beliefs about deception in children, adolescents and adults. Legal and Criminological Psychology, 11, 297-312.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004a). Let me inform you how to tell a convincing story: CBCA and reality monitoring scores as a function of age, coaching, and deception. Canadian Journal of Behavioral Science Revue, 36, 113-126.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004b). Detecting deceit via analyses of verbal and nonverbal behavior in children and adults. Human Communication Research, 30, 8-41.

Vrij, A., Edward, K., & Bull, R. (2001). People's insight into their own behaviour and speech content while lying. British Journal of Psychology, 92, 373-389.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. Journal of Nonverbal Behavior, 24, 239-263.

Vrij, A., & Mann, S. (2001). Telling and detecting lies in a high-stake situation: The case of a convicted murderer. Applied Cognitive Psychology, 15, 187-203.

Vrij, A., Mann, S., Robbins, E., & Robinson, M. (2006). Police officers' ability to detect deception in high stakes situations and in repeated lie detection tests. Applied Cognitive Psychology, 20, 741-755.

Vrij, A., & Taylor, R. (2003). Police officers' and students' beliefs about telling and detecting trivial and serious lies. International Journal of Police Science and Management, 5, 1-9.

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.

Zuckerman, M., Koestner, R., & Driver, R. (1981). Beliefs about cues associated with deception. Journal of Nonverbal Behavior, 6, 105-114.

Anhang 2a
Versuchsdesign und Szenarien für die verschiedenen Untersuchungsbedingungen

Situation	Ohne Vorbereitung	Mit Vorbereitung
Verspätung	Frau K. ist mit ihrem Freund verabredet. Sie sieht sich einen Spielfilm an und hat darüber die Zeit vergessen. Als sie auf die Uhr schaut, merkt sie, dass sie direkt losfahren müsste, um pünktlich am vereinbarten Treffpunkt zu sein. Doch da sie sich dazu entschließt, den Film noch zu Ende zu sehen, kommt sie erst 20 Minuten später los. Als sie gerade das Haus verlässt, trifft sie völlig überrascht auf ihren Freund. Er ist ihr schon entgegengekommen und scheint verärgert zu sein. Da sie befürchtet, dass es zum Streit kommen könnte, belügt sie ihn. Sie gibt fälschlicherweise an durch einen wichtigen beruflichen Anruf aufgehalten worden zu sein und es daher nicht früher geschafft zu haben.	Frau K. ist mit ihrem Freund verabredet. Sie sieht sich einen Spielfilm an und hat darüber die Zeit vergessen. Als sie auf die Uhr schaut, merkt sie, dass sie direkt losfahren müsste, um pünktlich am vereinbarten Treffpunkt zu sein. Doch da sie sich dazu entschließt, den Film noch zu Ende zu sehen, kommt sie erst 20 Minuten später los. Da sie befürchtet, dass es zum Streit kommen könnte, überlegt sie sich gründlich, was sie ihm sagen wird. Sie hat unterwegs ausreichend Zeit ihre Antworten zu planen. Ihr Freund wartet bereits am vereinbarten Treffpunkt. Wie sie es bereits erwartet hat, scheint er verärgert zu sein. Sie belügt ihn indem sie behauptet, durch einen wichtigen beruflichen Anruf aufgehalten worden zu sein und es daher nicht früher geschafft zu haben.

Anhang 2a (Fortsetzung)

Situation	Ohne Vorbereitung	Mit Vorbereitung
Affäre	Frau K. und Herr K. sind seit mehreren Jahren verheiratet. Seit kurzem hat Frau K. jedoch einen Geliebten, von dem ihr Mann nichts weiß. Als sie eines Abends nach Hause kommt, konfrontiert ihr Mann sie vollkommen unerwartet damit, dass er zufällig gesehen hat, wie sie am frühen Abend, obwohl sie angeblich arbeiten musste, in Begleitung eines attraktiven Mannes ein Restaurant betrat. Frau K. ist sehr überrascht, sie war tatsächlich mit ihrem Geliebten dort. Da sie befürchtet, dass ihr Mann sie verlassen könnte, wenn er von ihrer Affäre erfährt, belügt sie ihn. Sie gibt fälschlicherweise an, der Mann sei ein neuer Kunde und es habe sich um ein Geschäftsessen gehandelt.	Frau K. und Herr K. sind seit mehreren Jahren verheiratet. Seit kurzem hat Frau K. jedoch einen Geliebten, von dem ihr Mann nichts weiß. Ein Freund hat Frau K. gewarnt, dass ihr Mann zufällig gesehen hat, wie sie am frühen Abend, obwohl sie angeblich arbeiten musste, in Begleitung eines attraktiven Mannes ein Restaurant betrat. Sie war tatsächlich mit ihrem Geliebten dort. Da sie befürchtet, dass ihr Mann sie verlassen könnte, wenn er von ihrer Affäre erfährt, überlegt sie sich gründlich, was sie ihm dazu sagen wird. Sie hat ausreichend Zeit ihre Antworten zu planen. Wie sie es bereits erwartet hat, stellt ihr Mann sie zur Rede, als sie nach Hause kommt. Sie belügt ihn indem sie behauptet, dass der Mann ein neuer Kunde sei und es sich um ein Geschäftsessen gehandelt habe.

Anhang 2a (Fortsetzung)

Situation	Ohne Vorbereitung	Mit Vorbereitung
Totschlag	Nach dem Tod eines ihnen nahe stehenden Freundes kommt es zwischen den Eheleuten Frau und Herr K. zu einem Streit um den Nachlass. Der Streit eskaliert, die beiden brüllen sich an, bewerfen sich mit Gegenständen, werden handgreiflich, und im Affekt erschlägt Frau K. ihren Mann. Unmittelbar danach klingelt es überraschend und die wegen des Lärms von den Nachbarn alarmierte Polizei steht vor der Tür. Das Wohnzimmer ist durch den Streit sehr verwüstet. Da Frau K. befürchtet wegen des soeben begangenen Verbrechens inhaftiert zu werden, belügt sie die Polizei. Sie gibt fälschlicherweise an, es seien Einbrecher im Haus gewesen, die ihren Mann erschlagen hätten und danach geflohen seien.	Nach dem Tod eines ihnen nahe stehenden Freundes kommt es zwischen den Eheleuten Frau und Herr K. zu einem Streit um den Nachlass. Der Streit eskaliert, die beiden brüllen sich an, bewerfen sich mit Gegenständen, werden handgreiflich, und im Affekt erschlägt Frau K. ihren Mann. Danach geht sie erst einmal um den Block bis sie sich wieder gefangen hat. Da sie befürchtet, wegen des begangenen Verbrechens inhaftiert zu werden, überlegt sie sich daraufhin gründlich was sie der Polizei sagen wird. Sie hat ausreichend Zeit, ihre Antworten zu planen. Erst dann alarmiert sie die Polizei. Das Wohnzimmer ist durch den Streit sehr verwüstet. Sie belügt die Polizei indem sie angibt, es seien Einbrecher im Haus gewesen, die ihren Mann erschlagen hätten und danach geflohen seien.

Anhang 2a (Fortsetzung)

Situation	Ohne Vorbereitung	Mit Vorbereitung
Sexualdelikt	---	Frau A. und Frau K. wohnen im gleichen Haus und sind eng befreundet. Eines Tages wird Frau A. von einem gemeinsamen Nachbarn im Keller vergewaltigt. Aufgrund von unangenehmen Erfahrungen mit Gericht und Polizei sieht sie sich jedoch nicht dazu in der Lage, in dieser Sache auszusagen. Sie berichtet ihrer Freundin Frau K. von dem Vorfall. Aus Wut über diese Ungerechtigkeit und aus Angst, selbst möglicherweise Opfer dieses Nachbarn zu werden, entschließt sich Frau K. den Täter anzuzeigen. Um ihrer Freundin eine Aussage zu ersparen, behauptet Frau K. fälschlicherweise, sie selbst sei das Opfer dieser Vergewaltigung gewesen. Frau K. weiss, dass eine Aufdeckung ihrer Falschaussage erhebliche rechtliche Konsequenzen hätte. Daher überlegt sie sich gründlich, was sie der Polizei sagen wird. Sie hat ausreichend Zeit, um Ihre Aussage zu planen.
KG	Wenn andere Personen lügen, d.h. bewusst falsch aussagen, um jemanden zu täuschen, inwiefern unterscheidet sich dann deren Aussage von einer wahren Darstellung?	

Anm: N = 40 je Untersuchungsbedingung. --- Eine Bedingung mit kurzer Vorbereitung wurde für das Sexualdelikt nicht realisiert. Die Bezeichnungen Frau K. und Herr K. wurden außer für das Sexualdelikt in der Hälfte der Szenarien vertauscht.

STUDIE 3

Inter-Rater-Reliabilität und Validität der Aberdeen Report Judgment Scales

Die rechtspsychologische Forschung verfolgt unter anderem die Frage, wodurch sich wahre von erfundenen Aussagen unterscheiden lassen. Neben psychophysiologischen wurden non- und paraverbale Täuschungskorrelate sowie inhaltliche Aussagemerkmale untersucht. Untersuchungen mittels des sogenannten „Lügendetektors“ oder „Polygraphen“ stellen eine Kombination psychologischer Fragetechniken und physiologischer Messungen dar. Es wird angenommen, dass sich wahr aussagende und lügende Personen in ihrer physiologischen Reaktion auf spezifische Fragen hin unterscheiden. Diese Reaktionen werden beispielsweise über die Messung der Hautleitfähigkeit, Atemfrequenz und Herzrate abgebildet. Der bisherige Forschungsstand verweist zwar auf eine gute Validität des Polygraphen, zumindest für die sogenannte Tatwissenstechnik (vgl. Ben-Shakhar & Elaad, 2003, für einen metaanalytischen Forschungsüberblick), allerdings lässt sich diese durch gezielte Manipulationen deutlich mindern (z.B. Ben-Shakhar & Dolev, 1996; Elaad & Ben-Shakhar, 1991; Honts, Devitt, Winbush & Kircher, 1996; Honts, Raskin & Kircher, 1994). Unter anderem deshalb sind polygraphische Untersuchungen im deutschen Strafverfahren bislang nicht zugelassen (BGH 1 StR 156/98 - Urteil vom 17.12.1998, Landesgericht Mannheim). Des Weiteren wurde überpüft, ob es in Abhängigkeit vom Wahrheitsstatus Unterschiede im sichtbaren oder akustisch wahrnehmbaren Verhalten gibt. Als nonverbale Korrelate wurden beispielsweise Mimik und Gestik, als paraverbale Stimmhöhe oder gefüllte und ungefüllte Pausen untersucht. Allerdings wurden kaum Unterschiede von praktischer Bedeutsamkeit im nonverbalen (DePaulo, Lindsay, Malone, Muhlenbruck, Charlton & Cooper, 2003; Sporer & Schwandt, 2007) und paraverbalen (DePaulo et al., 2003; Sporer &

Schwandt, 2006) Verhalten von wahr aussagenden und lügenden Personen nachgewiesen. Zusammenfassend konnten Anwender bislang nicht von der Nützlichkeit psychophysiologischer, non- und paraverbaler Untersuchungen zur Entdeckung von Täuschung überzeugt werden.

In der deutschen Begutachtungspraxis hat sich vor allem die inhaltsorientierte Glaubhaftigkeitsdiagnostik durchgesetzt. Dabei werden Aussagen auf bestimmte inhaltliche Merkmale hin untersucht. Es wurden verschiedene Inhaltskriterien vorgestellt, von denen angenommen wird, dass sie unterschiedlich häufig oder in unterschiedlicher Qualität bei wahren und erfundenen Aussagen vorzufinden sind. Gemäß ihrem Entwicklungshintergrund lassen sich diese Merkmale als forensische, Realitätsüberwachungs- und sozial-kognitive Kriterien bezeichnen (Sporer, 1997b).

Forensische Glaubhaftigkeitsmerkmale/CBCA

Die Criteria-Based Content Analysis (CBCA) umfasst die wohl bekanntesten inhaltlichen Aussagemerkmale. Diese wurden von Steller und Köhnken (1989) auf der Grundlage der praktischen Erfahrungen und Publikationen von Sachverständigen zusammengestellt und der internationalen Öffentlichkeit zugänglich gemacht. Wegen ihres praktischen Entwicklungshintergrundes werden sie im Folgenden auch als forensische Kriterien bezeichnet. Die CBCA ist Teil der übergeordneten Statement Validity Analysis (SVA). Die SVA verlangt die kognitiven Leistungsvoraussetzungen (Aussagetüchtigkeit), die Aussagequalität und die Rahmenbedingungen der Aussageentwicklung (Aussagezuverlässigkeit) zu analysieren. Die Begutachtung der Aussagequalität erfolgt anhand der 19 CBCA-Merkmale. Es wird angenommen, dass diese eher in erlebnisbasierten als in erfundenen Aussagen vorzufinden sind. Da ihr Vorhandensein die Hypothese eines persönlichen Erfahrungsbezugs stärkt, sind sie als Glaubhaftigkeitsmerkmale aufzufassen.

Seit den 50er Jahren wird die SVA überwiegend in Fällen des inkriminierten sexuellen Kindesmissbrauchs angewendet. Eine intensive wissenschaftliche Auseinandersetzung mit der CBCA begann in den 80er Jahren (vgl. Sporer, 1982; Steller, 1988; Undeutsch, 1967; Wegener, 1997, für historische Überblicksdarstellungen). In seinem Urteil vom 30. Juli 1999 legte der deutsche Bundesgerichtshof Qualitätsstandards für forensisch-psychologische Sachverständigengutachten fest (BGH 1 StR 618/98 - Urteil vom 30.07.1999, Landesgericht Ansbach). Darin wird unter anderem ausgeführt, dass die CBCA dazu geeignet ist, die Glaubhaftigkeitsbegutachtung zu unterstützen:

> “Diese sog. Realkennzeichen [CBCA-Merkmale] können als grundsätzlich empirisch überprüft angesehen werden. Zwar handelt es sich um Indikatoren mit jeweils für sich genommen nur geringer Validität, d.h. mit durchschnittlich nur wenig über dem Zufallsniveau liegender Bedeutung. Eine gutachterliche Schlußfolgerung kann aber eine beträchtlich höhere Aussagekraft und damit Indizwert für die Glaubhaftigkeit zu beurteilender Angaben erlangen, wenn sie aus der Gesamtheit aller Indikatoren abgeleitet wird. […] Dementsprechend lagen die mit Realkennzeichen in Forschungsvorhaben erzielten Ergebnisse regelmäßig deutlich über dem Zufallsniveau. Allerdings bestanden dabei teilweise nicht unerhebliche Fehlerspannen.” (BGH 1 StR 618/98 - Urteil vom 30.07.1999, Landesgericht Ansbach)

Dennoch wird bereits durch diese juristische Stellungnahme deutlich, dass die inhaltliche Aussageanalyse anhand der forensischen Glaubhaftigkeitsmerkmale noch problembehaftet ist. In der rechtspsychologischen Literatur wurde vor allem die unzureichende theoretische Fundierung der CBCA kritisiert (Köhnken, 1990; Sporer, 1997a; Steller, Wellershaus & Wolf, 1992). Bereits Steller und Köhnken (1989) gaben zu bedenken, dass die CBCA überwiegend auf praktischen Erfahrungen von Sachverständigen basiert. Um die Annahme zu begründen, dass die CBCA-Merkmale eher in erlebnisbasierten als

in erfundenen Aussagen vorzufinden sind, wird lediglich auf die Ausführungen Undeutschs verwiesen. Undeutsch argumentierte in den 60er Jahren: “Wer etwas erzählt, was er nicht in der Realität erlebt hat, spricht unvermeidlich davon, wie der Blinde von den Farben“ (Undeutsch, 1967, S. 167). In Analogie dazu schlussfolgerte er, dass sich erlebnisbasierte Aussagen qualitativ von erfundenen unterscheiden sollten. Diese als Undeutsch-Hypothese (Steller, 1989, S. 145) bekannt gewordene Annahme genügt allerdings nicht den Ansprüchen an eine theoretische Untermauerung. Sie spezifiziert weder warum noch wann Unterschiede zwischen erlebnisbasierten und erfundenen Aussagen zu erwarten sind (Sporer, 1997a).

Realitätsüberwachungskriterien

Den Kritikpunkt an den forensischen Kriterien aufgreifend bemühte man sich darum, inhaltliche Aussagemerkmale aus dem sogenannten Realitätsüberwachungsansatz (RÜ, Johnson & Raye, 1981; vgl. Johnson, 2006, für eine Zusammenfassung des empirischen Forschungsstandes) abzuleiten. Dieser gedächtnispsychologische Ansatz geht davon aus, dass sowohl extern als auch intern generierte Ereignisse überdauernde Gedächtnisspuren hinterlassen, die sich qualitativ unterscheiden. Es wird postuliert, dass Erinnerungen an extern generierte bzw. erlebnisbasierte Ereignisse mehr kontextuelle Attribute, sensorische und semantische Details aufweisen als Erinnerungen an intern generierte bzw. vorgestellte Ereignisse. Umgekehrt werden bei intern generierten Ereignissen mehr Informationen hinsichtlich kognitiver Operationen während der Enkodierung erwartet.

Der Begriff der Realitätsüberwachung bezieht sich auf den Prozess, über den eine Person eine Erinnerung einer externen oder internen Quelle zuschreibt, d.h. entscheidet, ob eine Erinnerung auf eigenen Erfahrungen oder lediglich auf Vorstellungen basiert. Dieser Metagedächtnisprozess wird neben den

angewandten Entscheidungsprozessen von den Charakteristiken der zu beurteilenden Gedächtnisspuren beeinflusst.

Bezogen auf die Entdeckung von Täuschung wird postuliert, dass Beobachter anhand derselben Charakteristiken feststellen können, ob die Aussage einer anderen Person erlebnis- oder vorstellungsbasiert ist. Die Vorhersagen des RÜ-Ansatzes werden damit auf die interpersonelle Ebene (Johnson, Bush & Mitchell, 1998) bzw. die Meta-Meta-Ebene (Sporer, 2004) übertragen. So wäre gemäß dem RÜ-Ansatz zu erwarten, dass Beobachter bei wahren Aussagen mehr kontextuelle, sensorische und semantische, hingegen weniger kognitive Informationen vorfinden als bei erfundenen. Im Gegensatz zu den forensischen CBCA-Kriterien sprechen also nicht alle aus dem RÜ-Ansatz abgeleiteten Merkmale für die Glaubhaftigkeit einer Aussage. Vielmehr werden Verweise auf kognitive Operationen als Lügenindikator aufgefasst.

Forensische und Realitätsüberwachungs-Kriterien im Vergleich

Die CBCA- und RÜ-Kriterien wurden hinsichtlich ihrer Gemeinsamkeiten und Unterschiede untersucht. Dazu führte Sporer verschiedene Laborstudien durch, in denen Rater dasselbe Stimulusmaterial anhand von elf CBCA- und acht RÜ-Merkmalen beurteilten (Sporer, 1997a, 1997b; Sporer & Bursch, 1996). Die RÜ-Merkmale waren zuvor empirisch aus einer modifizierten Version des Memory Characteristics Questionnaires von Johnson, Foley, Suengas und Raye (1988) abgeleitet worden (Sporer & Küpper, 1995, vgl. auch Sporer & Küpper, 2004). Als Untersuchungsmaterial wurden die transkribierten Aussagen von Erwachsenen über persönlich bedeutsame Ereignisse verwendet, die entweder erlebnisbasiert oder erfunden waren. Wiederholt ließen sich substanzielle Überlappungen zwischen den CBCA- und RÜ-Kriterien nachweisen, die es erlaubten, fünf bis sechs gemeinsame Faktoren zu extrahieren (Sporer, 1997a, 1997b; Sporer & Bursch, 1996).

Trotz dieser Gemeinsamkeiten führten die meisten Autoren getrennte Analysen durch, um die Klassifikationsgüte der forensischen und RÜ-Kriterien einander vergleichend gegenüberzustellen (Granhag, Strömwall & Landström, 2006; Vrij, Akehurst, Soukara & Bull, 2004a, 2004b; Vrij, Edward & Bull, 2001b; Vrij, Edward, Roberts & Bull, 2000). Durch die Kombination der CBCA- und RÜ-Kriterien ließ sich jedoch eine höhere Klassifikationsgüte erzielen als bei deren getrennter Analyse (Sporer, 1997a; Strömwall, Bengtsson, Leander & Granhag, 2004).

Die inkrementelle Validität verweist auf die Bedeutsamkeit der verbleibenden Unterschiede zwischen den CBCA- und RÜ-Merkmalen. Die RÜ-Kriterien beziehen sich ausschließlich auf Wahrnehmungs- und kognitive Merkmale im weitesten Sinne. Die CBCA wiederum berücksichtigt auch soziale Aspekte und umfasst beispielsweise die Schilderung von Interaktionen als Qualitätsmerkmal erlebnisbasierter Aussagen. Entsprechend zeigten die faktorenanalytischen Untersuchungen von Sporer (1997a, 1997b), dass dieses forensische Merkmal einen Faktor markiert, auf dem die RÜ-Kriterien nicht bedeutsam laden (Sporer 1997b: alle Faktorladungen $\leq .176$; Sporer 1997a: alle Faktorladungen $\leq .422$). Doch auch aus dem RÜ-Ansatz wurden Kriterien abgeleitet, die bei den forensischen Glaubhaftigkeitsmerkmalen nicht explizit berücksichtigt werden. Dies gilt beispielsweise für das Glaubhaftigkeitsmerkmal sensorische Details (Interkorrelationen zum global gefassten CBCA-Merkmal Quantität von Details $.19 \geq \underline{r} \geq .26$; Sporer 1997a, 1997b; Sporer & Bursch, 1996). Zusammenfassend ließen die dargestellten Befunde eine Integration der CBCA- und RÜ-Kriterien sinnvoll erscheinen.

Sozial-Kognitive Kriterien/ARJS

Sporer wiederum stellte sozial-kognitive Kriterien vor, die sogenannten Aberdeen Report Judgment Scales (ARJS; Sporer, 1996/1998/2004). Diese entwickelte er sowohl theoriegeleitet als auch auf der Grundlage seiner zuvor

dargestellten faktorenanalytischen Befunde (Sporer, 1997a, 1997b; Sporer & Bursch, 1996). So wurden die CBCA- und RÜ-Kriterien integriert und vor dem Hintergrund gedächtnis- und sozialpsychologischer Überlegungen ergänzt und modifiziert. Zudem wurden sämtliche Kriterien so formuliert, dass sie potenziell einen breiten Anwendungsbereich der ARJS ermöglichen. Sporer (2004) weist jedoch darauf hin, dass beim derzeitigen Forschungsstand noch unklar ist, ob die ARJS diese Zielsetzung erfüllen:

> “Im Gegensatz zu bisherigen Ansätzen, die sich lediglich auf Anwendungen in spezifischen Bereichen konzentriert haben, viz. SVA/CBCA auf Fälle des inkriminierten sexuellen Kindesmissbrauchs, oder RÜ auf suggerierte/induzierte Erinnerungen, wurden die ARJS so konzipiert, dass sie Anwendungen in verschiedenen Bereichen erlauben, sowohl bei Erwachsenen als auch bei Kindern. Ob die Validität dieses umfassenden Ansatzes für verschiedene Bereiche und Populationen demonstriert werden kann oder nicht, ist eine empirische Frage.” (Sporer, 2004, S. 92, Übersetzung der Verfasserin)

Die 52 ARJS-Kriterien sind in Tabelle 3.1 mit Verweisen auf ihre theoretische Fundierung aufgeführt. Die theoretische Fundierung der einzelnen ARJS-Skalen wird später im Rahmen der Diskussion genauer erläutert. Die nachfolgenden Ausführungen konzentrieren sich darauf, die Unterschiede zwischen den ARJS einerseits und den forensischen und Realitätsüberwachungskriterien andererseits herauszuarbeiten.

Für die ARJS wurden einzelne CBCA-Kriterien vor dem Hintergrund des RÜ-Ansatzes umformuliert und spezifiziert. So erfordern die ARJS beispielsweise keine globale Beurteilung des CBCA-Kriteriums kontextuelle Einbettung, sondern erfassen räumliche und zeitliche Details getrennt.

Tabelle 3.1
ARJS-Skalen und Items sowie deren Entwicklungshintergrund

ARJS-Skalen und Items	CBCA	RÜ	ABM	Soz.
Skala 1: Realismus und logische Struktur				
(01) Unplausible Details *	+			+
(02) Realismus		+		
(03) Widersprüche	+			+
(04) Rekonstruierbare Struktur		+		
Skala 2: Klarheit und Lebendigkeit				
(05) Klarheit		+	+	
(06) Visuelle Details		+	+	
(07) Lebendigkeit		+	+	
(08) Spontane Organisation	+			
Skala 3: Details				
(09) Details im Hauptereignis	+	+	+	
(10) Details im Nebenereignis		?		
(11) Präzise Details		?		
(12) Überflüssige Details	+			-
Skala 4: Räumliche Details				
(13) Räumliche Details gesamt		+	+	
(14) Details zu Ort und Umgebung		+	+	
(15) Anordnung von Gegenständen		+		
(16) Anordnung von Personen		+		
Skala 5: Zeitliche Details	+			
(17) Zeitliche Details gesamt		+	+	
(18) Angaben zur Jahreszeit		+	+	
(19) Angaben zum Jahr		+	+	
(20) Angaben zu Tag oder Datum		+	+	
(21) Angaben zur Tages- oder Uhrzeit		+	+	
Skala 6: Sinneseindrücke				
(22) Geräusche		+		
(23) Gerüche	+	+		
(24) Tastempfindungen		+		
(25) Geschmacksempfindungen		+		

Tabelle 3.1 (Fortsetzung)

ARJS-Skalen und Items	CBCA	RÜ	ABM	Soz.
Skala 7: Emotionen und Gefühle				
(26) Emotionen und Gefühle	+	+	+	
(27) Intensive Gefühle		+	+	
(28) Emotionstypen	+			
(29) Gefühlsverlauf	+			
Skala 8: Gedanken				
(30) Gedanken	+	-		
(31) Präzise Gedanken		-		
(32) Schlussfolgernde Prozesse		-		
Skala 9: Memorieren und Gedächtnis				
(34) Wiederholtes Nachdenken		+	+	
(35) Wiederholtes Erzählen		+	+	
(36) Ereignisse vorher	+	+		
(37) Ereignisse nachher	+	+		
(38) Unterstützende Erinnerungen			+	
Skala 10: Nonverbale und verbale Interaktionen				
(39) Nonverbale Interaktionen	+			
(40) Präzise nonverbale Interaktionen	+			
(41) Verbale Interaktionen	+			
(42) Präzise verbale Interaktionen	+			
(43) Gedanken und Gefühle anderer Personen	+			
Skala 11: Komplikationen und ungewöhnliche Details				
(44) Komplikationen	+		+	
(45) Ungewöhnliche Details	+		+	
Skala 12: Fehler und sozial Unerwünschtes				
(33) Sicherheit der eigenen Angaben *		-		+
(46) Korrekturen oder Präzisierungen	+			+
(47) Erinnerungslücken	+	-		+
(48) Mangel an sozialer Erwünschtheit				+

Tabelle 3.1 (Fortsetzung)

ARJS-Skalen und Items	CBCA	RÜ	ABM	Soz.
Skala 13: Persönliche Signifikanz				
(49) Persönliche Bedeutung			+	
(50) Scheinbar ernsthafte Folgen		+		
(51) Tatsächlich ernsthafte Folgen		+		
(52) Persönlichkeit des Erzählers		+		

Anm. ABM = Arbeitsgedächtnismodell, Soz. = Sozialpsychologische Theorien und Befunde. * Die Items werden eher in erfundenen als in wahren Aussagen erwartet und daher später umtransformiert. Die Klammern indizieren, dass die entsprechenden ARJS-Merkmale durch die CBCA nur global erfasst werden. + = Item wurde auf der Grundlage der entsprechenden Kriterien, Befunde oder Theorien abgeleitet; - = Das Item wird in den entsprechenden Kriterien, Befunden oder Theorien im Gegensatz zu den ARJS als Lügenmerkmal aufgefasst.

Allerdings reicht der rein kognitionspsychologische RÜ-Ansatz nicht aus, um die inhaltlichen Aussagemerkmale theoretisch zu untermauern. Daher gingen in die Entwicklung der ARJS auch sozialpsychologische Theorien und Forschungsbefunde mit ein. Beispielsweise argumentiert Sporer (1998) auf der Grundlage des Impression-Management-Ansatzes (DePaulo & Friedman, 1998; DePaulo et al., 2003; Fiedler, 1989a, 1989b, Pontari & Schlenker, 2000; Schlenker & Weigold, 1992), dass Lügner um eine kompetente und moralisch einwandfreie Selbstdarstellung bemüht sind. Unter der Annahme, dass Fehler und sozial Unerwünschtes mit Täuschung assoziiert werden (vgl. Studie 2), sollten Lügner entsprechendes Verhalten gezielt vermeiden. Infolgedessen wären bei erfundenen Aussagen weniger Fehler und sozial Unerwünschtes zu erwarten als bei wahren.

Des Weiteren wurden Theorien und Forschungbefunde zum autobiographischen Gedächtnis bei der Konzeption der ARJS berücksichtigt (z.B. Anderson & Conway, 1997; Brewer, 1986, 1996; Conway, 1990; Larsen, 1998). So wurden Unterschiede in der Erinnerung an zentrale und an periphere Details festgestellt (vgl. Christianson, 1992, für einen Überblick). Zentrale Details, die sich auf das Kernereignis beziehen, werden bei emotionalen Ereignissen besser erinnert als bei neutralen. Hingegen werden periphere Details, die irrelevante oder nebensächliche Informationen darstellen, bei neutralen Ereignissen besser erinnert als bei emotionalen. Für die Glaubhaftigkeitsanalyse von Aussagen zu neutralen Ereignissen (z.B. Alibis) könnten demnach auch periphere Details nützlich sein. Daher erfordern die ARJS eine getrennte Beurteilung von zentralen und peripheren Details bzw. von Details, die auf das Haupt- und auf Nebenereignisse bezogen sind.

Schließlich ist darauf hinzuweisen, dass sich die Vorhersagen der ARJS und des RÜ-Ansatzes zuweilen widersprechen. Die unter der Skala Gedanken zusammengefassten Aussagemerkmale werden bei den ARJS als Wahrheitskriterien aufgefasst. Der RÜ-Ansatz postuliert im Gegensatz dazu, dass

kognitive Operationen als Lügenkriterium zu werten sind. Allerdings sind unterschiedliche Operationalisierungen dieses Merkmals bei den ARJS und der Forschung zum RÜ-Ansatz festzustellen.

Im Rahmen der Forschung zum RÜ-Ansatz wurden kognitive Operationen sehr unterschiedlich definiert. So verwiesen Johnson et al. (1998) als Beispiele für kognitive Operationen auf “Prozesse beim Sehen eines Objektes, sich die Ansicht eines Objektes vorzustellen, oder sich an frühere Erfahrungen mit einem Objekt zu erinnern“ (S. 200). Die Arbeitsgruppe um Vrij (Vrij et al., 2000; Vrij et al., 2004a, 2004b) wertete Aussagen wie „Sie war recht groß für ein Mädchen“ (Vrij et al., 2004a, S. 120; 2004b, S. 20, Übersetzung der Verfasserin) als Ausdruck kognitiver Operationen. Diese kognitiven Operationen beziehen sich auf Denkprozesse bzw. Attributionen während des Erzählens einer Geschichte. Im Gegensatz dazu bezieht sich der RÜ-Ansatz nur auf Prozesse zum Zeitpunkt des Ereignisses. Alonso-Quecuty (1992) wiederum sprach von „idiosynkratischen Informationen“ (S. 328), ohne deren Operationalisierung zu erläutern.

Es gab jedoch auch Bemühungen, Realitätsüberwachungskriterien auf der Grundlage eines standardisierten Fragebogens zu definieren. Johnson et al. (1988) entwickelten einen Fragebogen zur Selbstbeschreibung von Erinnerungsqualitäten. Suengas und Johnson (1988) sowie Sporer und Küpper (1995, 2004) analysierten dessen Faktorenstruktur und fassten die 38 Items zu fünf bzw. acht Skalen zusammen. Die differenzierte Faktorenstruktur von Sporer und Küpper ergab sich unter anderem daraus, dass kognitive Operationen im Gegensatz zu den Analysen von Suengas und Johnson einem eigenständigen Faktor zugeordnet wurden. Sporer und Küpper fassten fünf Items zur Skala kognitive Operationen zusammen. Diese Items indizierten eine gute Erinnerung an Gedanken während des Ereignisses sowie an Ereignisse, die zuvor und danach stattfanden. Auch Selbstbeschreibungen, wiederholt über das Ereignis nachgedacht und gesprochen zu haben, wurden der RÜ-Skala kognitive Operationen zugeordnet.

Bei den ARJS finden sich ähnliche Aussagemerkmale unter den beiden Skalen Memorieren und Gedächtnis sowie Gedanken. Die ARJS-Skala Memorieren und Gedächtnis umfasst die Schilderungen von Ereignissen vor oder nach dem Hauptereignis. In Anlehnung an das CBCA-Kriterium Raum-zeitliche Verknüpfungen werden diese Merkmale eher bei wahren als bei erfundenen Aussagen erwartet (vgl. Johnson, 1985). Zur Entdeckung von Täuschung werden in der Regel Aussagen zu autobiographischen Ereignissen untersucht, die längst vergangen sind. Wiederholt über ein selbst-erlebtes Ereignis nachzudenken und darüber zu sprechen sollte die Erinnerung daran festigen (Brown & Schopflocher, 1998). Ebenso kann die Erinnerung an ein Ereignis durch andere Ereignisse, die zur gleichen Zeit stattfanden, unterstützt werden (Johnson, 1985; Johnson et al., 1988). Daher stärken entsprechende Angaben die Hypothese einer wahrheitsgemäßen Schilderung.

Die ARJS-Skala Gedanken bezieht sich auf die Quantität und Qualität von Gedanken zum Zeitpunkt des geschilderten Ereignisses, sowie auf schlussfolgernde Prozesse. Die Schilderung von Gedanken wird in Übereinstimmung mit der CBCA, jedoch im Gegensatz zum RÜ-Ansatz, als Wahrheitskriterium aufgefasst. Schlussfolgernde Prozesse setzen allgemeines, personen-, und/oder situationsspezifisches Wissen voraus. Die Rekonstruktion längst vergangener Ereignisse könnte dadurch erleichtert werden und infolgedessen auch Erinnerungen an selbst-erlebte Ereignisse unterstützen.

Zusammenfassend wird im Rahmen der ARJS davon ausgegangen, dass Gedanken, Memorieren und Gedächtnisprozesse dazu beitragen, die Erinnerung an vergangene Ereignisse zu festigen. Infolgedessen sollten vor allem bei wahrheitsgemäßen Aussagen entsprechende Merkmale vorzufinden sein.

Forschungsstand zur Inter-Rater-Reliabilität von Glaubhaftigkeitsmerkmalen

Die CBCA-, RÜ- und ARJS-Kriterien sind nicht als psychometrische Testverfahren aufzufassen. Sie dienen vielmehr dazu den Prozess der qualitativen

Aussageanalyse, als Teil der übergeordneten Glaubhaftigkeitsbegutachtung, zu unterstützen. Dennoch ist auch für die Beurteilung inhaltlicher Aussagemerkmale ein zufrieden stellendes Maß an Auswertungsobjektivität zu fordern. Wenn verschiedene Rater in ihrem Verständnis der einzelnen Kriterien divergieren und deren Vorhandensein unterschiedlich einschätzen, kann die inhaltliche Aussageanalyse nicht valide sein. Zudem ist es auch für den Anwendungskontext wesentlich, dass Sachverständige ihre Urteilsbildung für das Gericht nachvollziehbar darstellen können (Küpper & Sporer, 1995).

Inter-Rater-Reliabilität der CBCA

Die Inter-Rater-Reliabilität der CBCA war bisher nur in drei veröffentlichten Studien zentraler Untersuchungsgegenstand (Anson, Golding & Gully, 1993; Gödert, Gamer, Rill & Vossel, 2005; Horowitz, Lamb, Esplin, Boychuk, Krispin & Reiter-Lavery, 1997). Doch auch in anderen Studien wurde das Stimulusmaterial, oder zumindest Teile dessen, von mehreren unabhängigen Ratern beurteilt. Dabei wurden für dichotome Beurteilungen meist prozentuale Übereinstimmungen (Colwell, Hiscock & Memon, 2002; Vrij, Kneller & Mann, 2000) und für abgestufte Ratingskalen korrelative Zusammenhänge zwischen den Ratern berichtet (Akehurst, Köhnken, & Höfer, 2001; Porter & Yuille, 1996; Porter, Yuille & Lehman, 1999; Santtila, Roppola, Runtti & Niemi, 2000; Vrij, Edward et al., 2000; Vrij, Edward & Bull, 2001a; Vrij et al., 2004a). Andere Autoren wiederum stellten verschiedene Indizes einander vergleichend gegenüber und berechneten auch zufallskorrigierte Maße der Inter-Rater-Reliabilität (Buck, Warren, Betman & Brigham, 2002; Craig, Scheibe, Raskin, Kircher & Dodd, 1999; Granhag et al., 2006; Sporer 1997a, 1997b; Strömwall et al., 2004; Vrij et al. 2001a).

Vrij (2005) veröffentlichte einen qualitativen Forschungsüberblick zur Inter-Rater-Reliabilität der CBCA. Dabei zählte er für jedes Kriterium aus, in wie vielen der von ihm recherchierten Studien gute Reliabilitäten erzielt wurden. Unabhängig vom spezifischen Reliabilitätsindex interpretierte er Werte ≥ .60 als Hinweis auf

eine gute Inter-Rater-Reliabilität. Er schlussfolgerte, dass die meisten CBCA-Kriterien in der Mehrzahl der Untersuchungen gute Interrater-Reliabilitäten aufwiesen. Für die beiden Merkmale unstrukturierte Produktion und spontane Verbesserungen wurde hingegen nur selten eine gute Reliabilität erzielt. Einschränkend ist jedoch zu beachten, dass es methodisch bedenklich ist, auf verschiedene Reliabilitätsindizes dieselben Interpretationsrichtlinien anzuwenden. So demonstrierten mehrere Studien, dass verschiedene Reliabilitätsindizes für dasselbe Datenmaterial unterschiedlich ausfallen können (z.B. Anson et al., 1993; Gödert et al., 2005; Horowitz et al., 1997; Sporer, 1997a).

Tabelle 3.2 fasst die Ergebnisse zur Inter-Rater-Reliabilität verschiedener Untersuchungen zusammen. Um die Vergleichbarkeit mit den vorliegenden Untersuchungsbefunden zu gewährleisten, wurden jedoch nur Studien aufgenommen, die korrelative Reliabilitätsmaße berichteten. Im Gegensatz zu Vrij (2005) wurde zudem auf unpublizierte Studien (Boychuk, 1991; Höfer, Akehurst & Metzger, 1996, beide zitiert nach Vrij, 2005) sowie auf eine niederländische Publikation (Winkel & Vrij, 1995, zitiert nach Vrij, 2005) verzichtet. Stattdessen wurden neuere Studien (Granhag et al., 2006; Strömwall et al., 2004, Vrij et al., 2004a) sowie deutschsprachige Untersuchungen (Sporer, 1997a, 1997b) ergänzt.

Zudem ist aufgeführt, wie viele Rater unabhängig voneinander Beurteilungen vornahmen. In den meisten Studien führten nur zwei Rater die Aussageanalyse durch und es wurden nie mehr als drei Rater eingesetzt. Des Weiteren ist Tabelle 3.2 zu entnehmen, welche Beurteilungsskala die Rater verwendeten und wie viele Aussagen sie beurteilten. Diese Angaben lassen erkennen, dass bei Häufigkeitsauszählungen der CBCA-Kriterien oftmals bessere Reliabilitäten erzielt wurden als bei der Verwendung abgestufter Beurteilungsskalen (vgl. auch Vrij, Evans, Akehurst & Mann, 2004).

Tabelle 3.2
Inter-Rater-Reliabilität der CBCA (Pearson-Korrelationen)

				CBCA-Kriterien																	
Studien	N	m	B	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Akehurst et al. (2001)	93	2	7	.34	.35	.68	.42	.44	.67	.49	.33	.55	-.04	-	.62	.58	.35	.02	-.06	-	-	-
Gödert et al. (2005)	102	3	4	.04	.38	.61	.50	.34	.83	.05	.24	.20	- .02	-	.57	.66	.28	.27	- .02	-	.29	-
Granhag et al. (2006)	80	2	3	.54	.61	.88	-	-	-	-	-	-	-	-	.95	-	.78	.92	.94	-	-	-
Granhag et al. (2006)	80	2	H	-	-	.98	-	.88	.95	-	.93	-	-	-	.92	-	.83	.89	.98	-	-	-
Porter & Yuille (1996)	60	2	3	.80	.80	-	-	-	-	-	-	.80		-	-	-	-	-	-	-	-	-
Porter & Yuille (1996)	60	2	H	-	-	.80	-	-	-	.80	.80	-	-	.80	.67	-	.80	.80	-	-		-
Porter et al. (1999)	25	3	7	-	-	-	.24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Porter et al. (1999)	25	3	H	-	-	.70	.24	-	-	-	-	-	-	-	-	-	-	.70		-		-
Santilla et al. (2000)	20	2	3	.63	.63	-	.63	.63	.87	.63	.63	.63	.63	.63	.63	.63	.63	-	-	-	-	-
Sporer (1997a)	80	2	3	.15	.24	.30	.28	.45	.63	.69	.53	.48			.52	.36	-	-	-	-		-
Sporer, (1997b)	200	2	3	.37	.26	.36	.23	.53	.66	.67	.47	.19	.40	.20	.46	.44	-	-	-	-		-
Strömwall et al. (2004)	18	2	H	-	-	.98	.69	.96	.19	-	-	-	-	-	-	-	-	.99	1	-	-	-
Strömwall et al. (2004)	18	2	3	1	-	.89	1	.92	.46	-	.72	-	1	.68	1	1	.73	1	.88	-	-	.76

Tabelle 3.2 (Fortsetzung)

				CBCA-Kriterien																	
Studien	N	m	B	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Vrij & Mann (2006)	80	2	H	.35	.47	.95	.90	.42	.97	.47	.64	.60	-	-	-	-	-	-	.44	.61	.49	-
Vrij et al., (2000)	73	2	2	.55	.65	.90	.85	.90	.97	-	.77	.69	-	-	.58	.71	.54	.89	.70	-	1	-
Vrij et al. (2001b)	83	2	2	.51	.69	.90	.82	.91	.87	-	.65	-	-	-		.64	.58	.83	.74	-	-	-
Vrij et al. (2004a)	196	2	H	-	-	.90	.93	.58	.62	.51	.17	-	-	-	.61	.21	.71	.63	.78	-	-	-
Vrij et al. (2004a)	196	2	5	.49	.08	.56	.76	.55	.52	.30	.05	-	-	-	.68	.20	.57	.66	.68	-	-	-
Vrij et al. (2004b)	180	2	H	.34	.20	.76	.51	.52	.83	.35	.09	.44	-	.11	.45	.42	.35	.74	.58	-	-	-
Mittleres r				.57	.48	.81	.68	.69	.83	.54	.58	.54	.64	.56	.74	.68	.62	.79	.70	.61	.82	.76

Anm. N = Anzahl der Aussagen anhand derer die Inter-Rater-Reliabilität bestimmt wurde; m = Anzahl der Rater, B = Beurteilung erfolgte über Häufigkeitsauszählungen (H) oder mehrstufige Ratingskalen (2-, 3-, oder 7-stufig). - = Das entsprechende Kriterium wurde nicht erfasst oder es wurde keine korrelative Inter-Rater-Reliabilität berichtet. CBCA-Kriterien: 1 = Logische Konsistenz, 2 = Unstrukturierte Produktion, 3 = Detailreichtum, 4 = Kontexteinbettung, 5 = Wiedergabe von Interaktionen, 6 = Wiedergabe von Gesprächen, 7 = Unerwartete Komplikationen, 8 = Ungewöhnliche Details, 9 = Nebensächliche Details, 10 = Unverstandenes, 11 = Indirekt handlungsbezogene Schilderungen, 12 = Schilderung eigener psychischer Vorgänge, 13 = Schilderung psychischer Vorgänge des Täters, 14 = Spontane Verbesserungen, 15 = Eingeständnis von Erinnerungslücken, 16 = Einwände gegen die Richtigkeit der eigenen Aussage, 17 = Selbstbelastungen, 18 = Entlastung des Angeschuldigten, 19 = Deliktspezifische Aussageelemente. Mittleres r = Die Korrelationen wurden Fisher Z-transformiert, dann wurden zunächst die Befunde derselben Untersuchung, danach über alle verschiedenen Studien hinweg ungewichtet gemittelt und zurück in r transformiert.

Die Beschränkung auf Studien, die dieselben Reliabilitätsindizes berichteten, eröffnet die Möglichkeit die bisherigen Befunde direkt zu vergleichen und zu integrieren. Allerdings indizieren gleiche Abstände zwischen zwei Korrelationen im Extrembereich stärkere Unterschiede in der Beurteilerübereinstimmung als im mittleren Bereich (Bortz, 1999). Um eine Gleichabständigkeit der Differenzen zu erzielen, wurden die berichteten Korrelationen zunächst Fisher's Z-transformiert. Danach wurden die Werte der einzelnen Studien gemittelt und in r zurücktransformiert. Die für jedes Merkmal resultierenden mittleren Reliabilitäten sind ebenfalls in Tabelle 3.2 aufgeführt.

Für individualdiagnostische Zwecke sind höhere korrelative Übereinstimmungen zwischen Raterurteilen erforderlich als für Gruppenvergleiche. Wirtz und Caspar (2002) demonstrierten anhand eines Rechenbeispiels den Zusammenhang zwischen der Reliabilität von Raterurteilen, dem Standardmessfehler und dem Vertrauensintervall. Sie schlussfolgerten, dass sich die individuelle Merkmalsausprägung einer Person erst bei Reliabilitäten $r \geq .85$ präzise einschätzen lässt. Hingegen forderten sie für Gruppenvergleiche, die beispielsweise zu Forschungszwecken durchgeführt werden, korrelative Übereinstimmungen von $r \geq .70$. Aus Tabelle 3.2 ist ersichtlich, dass die mittleren Korrelationen für elf der 19 CBCA-Kriterien diesen Richtwert erzielten. Dabei waren für die Kriterien Detailreichtum, Eingeständnis von Erinnerungslücken, Einwände gegen die Richtigkeit der eigenen Aussage in mindestens der Hälfte der Studien gute Reliabilitäten von $r \geq .70$ nachweisbar. Dies galt auch für das Kriterium deliktspezifische Aussageelemente, dessen Inter-Rater-Reliabilität jedoch nur in einer Studie berichtet wurde. Im Gegensatz dazu wurden für die übrigen Kriterien trotz guter mittlerer Reliabilität in den meisten Studien keine entsprechend hohen Korrelationen nachgewiesen.

Insgesamt weisen die bisherigen Forschungsbefunde darauf hin, dass sich einzelne CBCA-Kriterien reliabel erfassen lassen. Allerdings ist eine starke Variabilität in der Anzahl der untersuchten Merkmale festzustellen. Obwohl eine

Vielzahl an CBCA-Untersuchungen vorliegt, wurde die Inter-Rater-Reliabilität für einige Kriterien bislang kaum überprüft.

Inter-Rater-Reliabilität der RÜ-Kriterien

Zur Inter-Rater-Reliabilität der RÜ-Kriterien liegen vergleichsweise weniger Untersuchungen vor. Die bisherigen Forschungsbefunde sind Tabelle 3.3 zu entnehmen. Zwei Laboruntersuchungen, in denen Aussagen von Kindern anhand von RÜ-Kriterien beurteilt wurden, berichteten durchaus vielversprechende Befunde. Granhag et al. (2006) sowie Strömwall et al. (2004) ließen jeweils 20% des Stimulusmaterials (16 und 18 Transkripte, respektive) von zwei unabhängigen trainierten Ratern anhand von 3-stufigen Antwortskalen und über Häufigkeitsratings beurteilen. In beiden Studien ergaben sich gute Inter-Rater-Reliabilitäten für die meisten RÜ-Kriterien. Lediglich die Kriterien Rekonstruierbarkeit (Granhag et al., 2006) und Realitätsnähe (Strömwall et al., 2004) wurden nicht reliabel erfasst. Auch für Aussagen von Erwachsenen gibt es Hinweise auf eine gute Inter-Rater-Reliabilität der RÜ-Kriterien. So berichteten Vrij und seine Kollegen hohe Korrelationen bei Häufigkeitsauszählungen (Vrij, Edward & Bull, 2001a; Vrij, Edward et al., 2000; Vrij, Evans et al., 2004). Allerdings zeigten sich bei Vrij et al. (2004a) Urteilsdiskrepanzen zwischen trainierten Ratern bezüglich des RÜ-Kriteriums kognitive Operationen.

Bei der Verwendung von Ratingskalen wurden allerdings geringere Inter-Rater-Reliabilitäten erzielt (Küpper & Sporer, 1995; Sporer, 1997a; 1997b). Küpper und Sporer (1995, vgl. auch Sporer und Küpper, 1995) ließen 40 Aussagen von Erwachsenen hinsichtlich der Ausprägung von acht RÜ-Kriterien anhand 7-stufiger Ratingskalen beurteilen. Erneut erwies sich die Erfassung kognitiver Operationen als problematisch. Zudem wurden die RÜ-Kriterien Klarheit und Lebendigkeit, Rekonstruierbarkeit und Realitätsnähe nicht reliabel beurteilt. In zwei weiteren Studien von Sporer (1997a, 1997b) zeigten sich zwar hohe prozentuale Übereinstimmungen für diese drei Merkmale zwischen zwei Ratern, andere

Tabelle 3.3

Inter-Rater-Reliabilität der RÜ-Kriterien (Pearson-Korrelationen)

				RÜ-Kriterien												
Studien	N	m	Beurteilung	1	2	2a	2b	2c	2d	2e	3	4	5	6	7	8
Küpper & Sporer (1995)	40	3	7-stufig	.30	.77	-	-	-	-	-	.76	.86	.43	.61	.13	.34
Granhag et al. (2006)	80	2	Häufigkeiten	-	-	-	1	1	-	-	.84	.96	-	.93	.89	-
			3-stufig	.68	-	-	.82	1	-	-	.66	.83	.54	.94	.89	.85
Sporer (1997a)	80	2	3-stufig	.25	.16	-	-	-	-	-	.47	.43	.01	.64	.21	.36
Sporer (1997b)	200	2	3-stufig	.12	.41	-	-	-	-	-	.52	.47	.40	.38	.21	.44
Strömwall et al. (2004)	18	2	Häufigkeiten	-	-	.97	1	-	1	1	.93	.95	-	-	.90	-
			3-stufig	.71	-	.80	1	-	1	1	1	.79	.73	-	.90	.52
Vrij et al. (2000)	73	2	Häufigkeiten	-	-	.96	.77	-	-	-	.72	.85	-	-	.75	-
Vrij et al. (2001a)	73	2	Häufigkeiten	-	-	.76	.96	-	-	-	.68	.87	-	-	.61	-
Vrij et al. (2004a)	196	2	Häufigkeiten	-	-	.80	.92	-	-	-	.62	.78	-	-	.54	-
Mittleres r				.45	.49	.89	.95	.99	.99	.99	.75	.81	.45	.71	.62	.54

Anm. N = Anzahl der Aussagen, m = Anzahl der Rater, Beurteilung erfolgte über Häufigkeitsauszählungen oder mehrstufige Ratingskalen. - = Das entsprechende Kriterium wurde nicht erfasst oder es wurde keine korrelative Inter-Rater-Reliabilität berichtet. RÜ-Kriterien: 1 = Klarheit & Lebendigkeit, 2= Sensorische Informationen, 2a = Visuelle Details, 2b = Geräusche, 2c = Gerüche, 2d = Geschmacksempfindung, 2e = Tastempfindung, 3 = Räumliche Informationen, 4 = Zeitliche Informationen, 5 = Rekonstruierbarkeit, 6 = Gefühle, 7 = Kognitive Operationen, 8 = Realitätsnähe. Mittleres r = Die Korrelationen wurden Fisher Z-transformiert, verschiedene Werte derselben Studie zusammengefasst und dann ungewichtet über alle Studien gemittelt und zurück in r transformiert.

Reliabilitätsmaße fielen jedoch gering aus. Der Autor führte die geringe Reliabilität sowohl auf die Globalität der Urteile als auch auf Boden- und Deckeneffekte zurück. Über alle Studien gemittelt wurden für neun Kriterien Reliabilitäten von $\underline{r} \geq .70$ nachgewiesen. Allerdings wurden drei dieser Merkmale, Gerüche, Geschmacks- und Tastempfindungen, nur selten untersucht.

Zusammenfassend berichten die wenigen bisherigen Untersuchungen widersprüchliche Befunde zur Inter-Rater-Reliabilität der RÜ-Kriterien. Unterschiede im verwendeten Stimulusmaterial sowie die fehlende Standardisierung bei der Verwendung von Beurteilungsskalen tragen möglicherweise dazu bei. Häufigkeitsauszählungen führten im Allgemeinen zu höheren Reliabilitäten. Dies könnte jedoch ein statistisches Artefakt sein, insofern als die Verteilungen oft schief gewesen sein dürften.

Inter-Rater-Reliabilität der ARJS

Im Gegensatz zu den CBCA- und RÜ-Kriterien werden die ARJS standardisiert anhand 7-stufiger Skalen beurteilt. Zur Interrater-Reliabilität der ARJS liegen bislang fünf unabhängige Studien vor (Barnier, Sharman, McKay & Sporer, 2005; Sporer & Burghardt, 2004; Sporer, 1998; Sporer, Bursch, Schreiber, Weiss, Hoefer, Sievers & Köhnken, 2000; Sporer & Walther, 2006). Die Befunde sind in Tabelle 3.4 zusammenfassend dargestellt.

Auf Skalenebene war die Inter-Rater-Reliabilität der sozial-kognitiven Merkmale meist zufrieden stellend bis gut. Für die Skalen Details, zeitliche Details sowie für nonverbale und verbale Interaktionen ergaben sich mittlere Reliabilitäten von $\underline{r} \geq .70$. Zudem zeigten sich für die Skalen Sinneseindrücke, Emotionen und Gefühle, Gedanken sowie für Memorieren und Gedächtnis Korrelationen von $\underline{r} \geq .65$. Die guten Reliabilitäten für diese sieben Skalen wurden in der Mehrzahl der Untersuchungen nachgewiesen.

Tabelle 3.4

Inter-Rater-Reliabilität der ARJS (Pearson-Korrelationen)

			ARJS-Skalen												
Studien	N	m	1	2	3	4	5	6	7	8	9	10	11	12	13
Barnier et al. (2005)	135	2	.63	.60	.56	.53	.71	.67	.65	.64	.53	.73	.53	.70	.69
Sporer (1998)	71	2	.03	.49	.69	.68	.79	.35	.72	.60	.71	.68	.32	.66	.43
Sporer & Burghardt (2004)	184	2	.30	.45	.68	.69	.87	.70	.56	.70	.62	.80	.19	.44	.39
Sporer et al. (2000/02)	160	3	.28	.55	.69	.54	.82	.71	.60	.61	.54	.58	.33	.43	.39
Sporer & Walther (2003)															
Freie Berichte	72	2	.55	.51	.69	.45	.94	.70	.81	.84	.75	.84	.71	.47	.10
Interviews	72	2	.41	.63	.77	.39	.92	.75	.75	.82	.78	.78	.61	.35	.15
Mittleres r			.36	.53	.67	.58	.84	.65	.67	.69	.64	.73	.42	.54	.42

Anm. N = Anzahl der Aussagen anhand derer die Inter-Rater-Reliabilität bestimmt wurde, m = Anzahl der Rater, Beurteilung erfolgte über 7-stufige Ratingskalen. - = Das entsprechende Kriterium wurde nicht erfasst oder es wurde keine korrelative Inter-Rater-Reliabilität berichtet. ARJS-Skalen: 1 = Realismus und logische Struktur, 2 = Klarheit und Lebendigkeit, 3 = Details, 4 = Räumliche Detaills, 5 = Zeitliche Details, 6 = Sinneseindrücke, 7 = Emotionen und Gefühle, 8 = Gedanken, 9 = Memorieren und Gedächtnis, 10 = Nonverbale und verbale Interaktionen, 11 = Komplikationen und ungewöhnliche Details, 12 = Fehler und sozial Unerwünschtes, 13 = Persönliche Signifikanz. Mittleres r = Die Korrelationen wurden Fisher Z-transformiert, dann wurden die beiden Bedingungen in der Untersuchung von Walther gemittelt und die resultierenden Werte mit denen der anderen Studien ungewichtet gemittelt und zurück in r transformiert.

Die reliable Erfassung der ARJS-Skala Gedanken steht im Kontrast zu der geringen Inter-Rater-Reliabilität für das RÜ-Kriterium kognitive Operationen. Dies verweist darauf, dass durch Standardisierung und präzisere Operationalisierung die Reliabilität der Beurteilungen verbessert wird. Für die Skalen Realismus und logische Struktur, Klarheit und Lebendigkeit, sowie persönliche Signifikanz wurden hingegen meist keine guten Inter-Rater-Reliabilitäten ermittelt. Dies ließ sich jedoch zumeist auf Boden- und Deckeneffekte zurückführen. Korrelationsmaße setzen eine deutliche Merkmalsvarianz im Datenmaterial voraus. Unterscheiden sich die zu beurteilenden Aussagen nur geringfügig in ihrer wahren Merkmalsausprägung, so ist es schwierig diese minimalen Varianzunterschiede aufzuklären. Die als Stimulusmaterial verwendeten Aussagen von Erwachsenen wurden beispielsweise in der Regel als sehr realistisch und klar beurteilt (Barnier et al., 2005; Sporer, 1998; Sporer & Walther, 2006). Infolgedessen war die Reliabilität der entsprechenden ARJS-Skala oftmals gering.

Insgesamt erscheinen die Befunde zur Inter-Rater-Reliabilität der ARJS weitestgehend konsistent. Die verbleibenden Variationen sind möglicherweise auf Unterschiede im Training und Erfahrungshintergrund der Rater zurückzuführen. So wurden beispielsweise in der Untersuchung von Sporer und Walther (2006) die Aussagen von zwei Ratern analysiert, die äußerst erfahren waren. Entsprechend wurden in dieser Untersuchung oftmals sehr gute Inter-Rater-Reliabilitäten erzielt.

Forschungsstand zur Validität von Glaubhaftigkeitsmerkmalen

Die Frage nach der Validität von Glaubhaftigkeitsmerkmalen hat im Vergleich zur Inter-Rater-Reliabilität wesentlich mehr Forschungsaufmerksamkeit erfahren. Eine Vielzahl von Laborstudien hat überprüft, ob sich wahre und erfundene Aussagen anhand inhaltlicher Merkmale unterscheiden lassen. Dabei wurden wahrheitsgemäße oder falsche Aussagen über ein Ereignis verwendet, in das die Probanden im Rahmen des Experiments involviert wurden oder das ihnen auf Video gezeigt wurde. In anderen Studien wiederum wurden die Probanden

dazu aufgefordert, persönliche Lebensereignisse zu schildern oder zu erfinden. Eines der Probleme, das vielen Laborstudien zur Entdeckung von Täuschung anhaftet, ist die Sanktionierung der Lüge durch den Versuchsleiter bzw. das Experiment selbst (vgl. Miller & Stiff, 1993). Durch die Instruktion zu lügen werden viele der im Feld wirksamen Begleiterscheinungen (z.B. moralische Verwerflichkeit, Gefahr der Lüge überführt zu werden) reduziert. Die ökologische Validität von Laborstudien ist daher begrenzt. Umgekehrt lässt sich im Rahmen von Felduntersuchungen nicht belegen, dass eine Aussage tatsächlich der Wahrheit entspricht oder erfunden wurde. Gerichtsurteile oder Geständnisse sind lediglich Indikatoren des objektiven Wahrheitsstatus und sind oftmals nicht unabhängig von dem Ergebnis der inhaltlichen Aussageanalyse (Köhnken & Wegener, 1985). Zudem weisen die wenigen publizierten Feldstudien (Craig et al., 1999; Lamb, Sternberg, Esplin & Hershkowitz, 1997; Parker & Brown, 2000; Rassin & van der Sleen, 2005) teilweise erhebliche methodische Mängel auf (vgl. Vrij, 2005). Die nachfolgenden Ausführungen repräsentieren daher vor allem die im Rahmen von Laboruntersuchungen gewonnenen Forschungsbefunde.

Validität der CBCA-Kriterien

Zur Validität der CBCA-Kriterien liegen sowohl metaanalytische Befunde (DePaulo et al., 2003) als auch eine qualitative Literaturübersicht vor (Vrij, 2005). DePaulo et al. (2003) untersuchten in ihrer Metaanalyse unter anderem die Validität von 16 CBCA-Kriterien. Sie fanden signifikante Mittelwertsunterschiede hinsichtlich der CBCA-Merkmale logische Struktur ($\underline{d}$ = 0.25, Anzahl unabhängiger Hypothesentests $\underline{k}$ = 6)[1], spontane Korrekturen ($\underline{d}$ = 0.29, $\underline{k}$ = 5) und

[1] DePaulo et al. (2003) berichteten negative Effektstärken, wenn die entsprechenden Merkmale seltener bei erfundenen als bei wahren Aussagen vorzufinden waren. Negative Effektstärken indizierten demnach, dass ein Wahrheitsindikator vorliegt. Um die Vergleichbarkeit der Befunde von DePaulo et

Eingeständnis von Erinnerungslücken (d = 0.42, k = 5). Diese Merkmale waren bei wahren Aussagen erwartungsgemäß stärker ausgeprägt als bei erfundenen. Indirekt handlungsbezogene Schilderungen waren hingegen häufiger bei erfundenen als bei wahren Aussagen vorzufinden (d = -0.35, k = 3). Allerdings berücksichtigten DePaulo et al. nur den Forschungsstand bis 1999 und stützen ihre Analysen auf eine recht geringe Datenbasis von ein bis sechs unabhängigen Untersuchungen. Zudem wurden die CBCA-Merkmale quantitativer Detailreichtum, Wiedergabe von Gesprächen und deliktspezifische Aussagemerkmale von der Analyse ausgeschlossen.

Einen aktuelleren qualitativen Forschungsüberblick zur Validität aller 19 CBCA-Kriterien lieferte Vrij (2005). Dazu erstellte er eine tabellarische Übersicht der Befunde von fünf Feld- und 17 Laborstudien. Für jedes Kriterium wurde angegeben, ob es häufiger, seltener oder gleichermaßen oft in wahren und erfundenen Aussagen vorgefunden wurde. Zudem wurde der Anteil der Untersuchungen ausgezählt, die eine erwartungsgemäß höhere Aussagequalität wahrer im Vergleich zu erfundenen Aussagen berichteten. Allerdings verdecken diese prozentualen Angaben, dass einzelne Kriterien bislang kaum untersucht wurden. Des Weiteren zitierte Vrij Befunde zur Validität von CBCA-Gesamtwerten. Die CBCA-Kriterien sollten jedoch untereinander nur gering korrelieren, weil sie sich auf sehr unterschiedliche Aussageaspekte beziehen. Vor diesem Hintergrund erscheint es problematisch, die Bewertungen einzelner Kriterien ungewichtet aufzusummieren (z.B. Akehurst et al., 2001; Craig et al., 1999; Raskin & Esplin, 1991; Vrij et al., 2004a, 2004b). Die nachfolgenden Ausführungen beschränken sich daher auf die Validität einzelner Kriterien.

---

al. mit den vorliegenden Untersuchungsbefunden zu erleichtern, wurden die Vorzeichen der zitierten Effektstärken ausgetauscht. Infolgedessen sind Wahrheitsindikatoren durch positive und Lügenindikatoren durch negative Effektstärken gekennzeichnet.

Vrij (2005) schlussfolgerte, dass insbesondere das CBCA-Merkmal Detailreichtum valide zu sein scheint. Ebenso haben sich die Kriterien Raum-zeitliche Verknüpfungen und Wiedergabe von Gesprächen meist als valide erwiesen. Zudem wurden oftmals erwartungsgemäße Unterschiede zwischen wahren und erfundenen Aussagen hinsichtlich des Merkmals unstrukturierte Darstellung berichtet. Allerdings ist darauf hinzuweisen, dass Rater häufig in ihrem Verständnis dieses Aussagemerkmals divergieren (z.B. Akehurst et al., 2001; Anson et al., 1993; Vrij et al., 2004a, 2004b). In mehr als der Hälfte der von Vrij zitierten Untersuchungen ergaben sich zudem erwartungsgemäße Befunde für die Kriterien logische Konsistenz, Wiedergabe von Interaktionen und Schilderung ausgefallener Einzelheiten. Die empirischen Validitätsbefunde für die übrigen Merkmale sind vergleichsweise schwach. Insbesondere die Merkmale Einwände gegen die Richtigkeit der eigenen Aussage und phänomengemäße Schilderung unverstandener Handlungselemente bildeten meist keine Unterschiede zwischen wahren und erfundenen Aussagen ab. Für Selbstbelastungen wiederum ergaben sich entweder keine Unterschiede zwischen wahren und erfundenen Aussagen, oder sie fanden sich sogar häufiger bei erfundenen Aussagen.

Bezieht man auch neuere und deutschsprachige Untersuchungen mit ein, die nicht von Vrij (2005) berücksichtigt wurden, ändert sich kaum etwas an diesen Schlussfolgerungen. Die Validität der Kriterien quantitativer Detailreichtum und Wiedergabe von Gesprächen wurde erneut bestätigt (Akehurst et al., 2004; Gödert et al., 2005; Granhag et al., 2006; Niehaus, 2003, vgl. auch Niehaus, 2001). Für die Merkmale unstrukturierte Darstellung (Granhag et al., 2006; Niehaus, 2001) und Raum-zeitliche Verknüpfungen (Gödert et al., 2005) wurden erwartungskonträre Befunde berichtet. Hingegen konnten für die Merkmale phänomengemäße Schilderung unverstandener Handlungselemente und Selbstabwertungen erneut keine Unterschiede zwischen wahren und erfundenen Aussagen nachgewiesen werden (Gödert et al., 2005; Niehaus, 2003). Eine der neueren Studien berichtete erwartungsgemäße Befunde hinsichtlich des CBCA-Merkmals Einwände gegen

die Richtigkeit der eigenen Aussage (Granhag et al., 2006). In den übrigen Studien wurden jedoch erneut keine Unterschiede zwischen wahren und erfundenen Aussagen für dieses Merkmal aufgedeckt (Akehurst et al., 2004; Gödert et al., 2005; Niehaus, 2003).

Es bleibt jedoch kritisch darauf hinzuweisen, dass die geschilderten Schlussfolgerungen zur Validität einzelner CBCA-Merkmale auf Häufigkeitsauszählungen beruhen. Die Literaturübersicht von Vrij (2005) bietet keinerlei Hinweise auf die Stärke und damit die praktische Bedeutsamkeit der Effekte.

Um die Nützlichkeit der CBCA-Merkmale für die Glaubhaftigkeitsdiagnostik zu überprüfen, wurden multiple Diskriminanz- und Regressionsanalysen durchgeführt. Vrij (2005) zitierte 14 Studien, bei denen überprüft wurde, wie gut sich der objektive Wahrheitsstatus durch die untersuchten CBCA-Kriterien vohersagen ließ. Die Befunde verweisen auf überzufällige Klassifikationsraten zwischen 67% und 89%.

Zusammenfassend liegt eine Vielzahl an Untersuchungen zur Validität der CBCA-Kriterien vor. Allerdings wurden in keiner Laborstudie alle Kriterien überprüft. Einzelne Kriterien wurden von den Analysen ausgeschlossen, weil sie nicht auf das Stimulusmaterial anwendbar zu sein schienen oder aufgrund methodischer Probleme (z.B. mangelnde Reliabilität). Dadurch liegen für einzelne Aussagemerkmale, wie für deliktspezifische Aussageelemente, Entlastung des Angeschuldigten und Selbstbelastungen, kaum Validitätsbefunde vor. Doch auch für Kriterien, die häufig untersucht wurden, ist es kaum möglich präzise Aussagen hinsichtlich ihrer Validität treffen. Das Auszählen erwartungskonformer Befunde liefert zwar eine grobe Orientierung, ermöglicht es jedoch nicht, die Stärke der Effekte zusammenfassend einzuschätzen. Dazu wäre es erforderlich, die Befunde der bislang vorliegenden Einzelstudien zu den CBCA-Merkmalen metaanalytisch zu integrieren. Die Vergleichbarkeit einzelner Studien wird jedoch dadurch erschwert, dass unterschiedliche Auswertungsmethoden verwendet wurden. Auf

Effektstärkenmaße bzw. auf vollständige Informationen, um diese abzuleiten, wurde häufig verzichtet.

Validität der RÜ-Kriterien

DePaulo et al. (2003) berichteten in ihrer Metaanalyse auch Befunde zur Validität von sechs Realitätsüberwachungskriterien. Allerdings basierten ihre Berechnungen nur auf ein bis vier unabhängigen Untersuchungen. Zudem ließen sich die Effektstärken oftmals nicht präzise aus den Primärstudien ableiten. Daher sind diese metaanalytischen Befunde zur Validität der RÜ-Kriterien wenig aufschlussreich.

Ein aktueller und umfassender qualitativer Forschungsüberblick zur Validität der RÜ-Kriterien wurde von Masip, Sporer, Garrido und Herrero (2005) veröffentlicht. Dabei wurden neben der englischsprachigen Literatur auch Publikationen und Tagungsbeiträge in deutscher, französischer und spanischer Sprache berücksichtigt. Für die RÜ-Kriterien Realismus, kontextuelle, zeitliche und räumliche Informationen bewerteten Masip et al. (2005) die vorliegenden Forschungsbefunde als ermutigend. Diese Merkmale wurden oftmals häufiger in wahren als in erfundenen oder verfälschten Aussagen vorgefunden. Hingegen ergaben sich für internale Informationen bzw. kognitive Prozesse oftmals ernüchternde Befunde. Für dieses, nach dem RÜ-Ansatz als Lügenkriterium aufzufassende inhaltliche Merkmal, wurden oftmals keine Unterschiede zwischen wahren und erfundenen Aussagen nachgewiesen. Widersprüchliche Forschungsbefunde wurden für sensorische Informationen berichtet. Masip et al. führen dies auf Unterschiede in der Operationalisierung des Aussagemerkmals sowie auf Bodeneffekte zurück.

Durch die gemeinsame Betrachtung mehrerer RÜ-Kriterien ist es möglich, wahre und erfundene Aussagen überzufällig richtig zu klassifizieren. Masip et al. (2005) zitierten sechs Studien, die über diskriminanz- oder regressionsanalytische Verfahren Klassifikationsraten zwischen 69% und 85% erzielten.

Schließlich ist gemäß den Ausführungen von Masip et al. (2005) zu bedenken, dass sich in der Forschung zum RÜ-Ansatz bislang keine einheitlichen Definitionen und Beurteilungsformate durchgesetzt haben. Ebenso tragen Unterschiede in den Untersuchungsdesigns, dem verwendeten Stimulusmaterial und der Manipulation potenzieller Moderatoren zur Variabilität der Untersuchungsbefunde bei (Masip et al., 2005).

Validität der CBCA- und RÜ-Kriterien im Vergleich

Einige Autoren überprüften die Validität der CBCA- und der RÜ-Kriterien anhand desselben Stimulusmaterials. Dabei wurden Aussagen von Kindern (Granhag et al., 2006; Strömwall et al., 2004), Erwachsenen (Sporer, 1997a, 1997b; Sporer & Bursch, 1996; Vrij, Edward et al., 2000) oder verschiedenen Altersgruppen (Vrij et al., 2004a, 2004b) analysiert. Es ließen sich varianzanalytisch in der Regel signifikante Haupteffekte des Wahrheitsstatus auf die inhaltlichen Aussagemerkmale nachweisen. Für erlebnisbasierte Aussagen zeigten sich erwartungsgemäß höhere Beurteilungen als für erfundene. Dies galt sowohl für die CBCA- als auch für die RÜ-Kriterien.

Allerdings waren in einzelnen Studien wahre Aussagen ausführlicher als erfundene (Vrij, Edward et al., 2000; Strömwall et al., 2004). Dadurch bleibt unklar, ob eine höhere inhaltliche Qualität auf den Wahrheitsstatus oder die Länge der Aussagen zurückzuführen ist. Dieses Problem ergibt sich vor allem bei Häufigkeitsauszählungen (Vrij, Edward et al., 2000; für RÜ-Kriterien). Ratingskalen gelten hingegen als robust gegenüber Variationen in der Aussagelänge (Granhag et al., 2006; Strömwall et al., 2004).

Zudem wurde die Klassifikationsgüte der CBCA- und RÜ-Kriterien verglichen. Dazu wurden getrennte Diskriminanzanalysen durchgeführt, bei denen die Indikatoren der inhaltlichen Aussagequalität als Prädiktoren und der objektive Wahrheitsstatus als Kriteriumsvariable verwendet wurden. Dabei gingen teilweise die einzelnen Kriterien (Sporer, 1997a, 1997b; Sporer & Bursch, 1996), teilweise

jedoch auch daraus abgeleitete Gesamtwerte (Granhag et al., 2006; Strömwall et al., 2004; Vrij, Edward et al., 2000; Vrij et al., 2004a, 2004b) als Prädiktorvariablen in die Analysen mit ein. Häufig ergab sich eine bessere Klassifikationsgüte für wahre als für erfundene Aussagen. Dies lässt sich vermutlich darauf zurückführen, dass fast alle inhaltlichen Aussagemerkmale als Wahrheitskriterien aufzufassen sind. Lediglich das RÜ-Kriterium kognitive Operationen gilt als Lügenkriterium. Es wurde jedoch teilweise von den Analysen ausgeschlossen (Vrij et al., 2004b; Vrij, Edward et al., 2000). Die RÜ-Kriterien erzielten meist eine etwas bessere Klassifikationsgüte als die forensischen Merkmale (Granhag et al., 2006; Sporer, 1997a, 1997b; Sporer & Bursch, 1996; Strömwall et al., 2004; Vrij et al., 2004a, 2004b). Anhand der CBCA-Merkmale ließen sich 54-73%, anhand der RÜ-Merkmale 62-74% des Stimulusmaterials richtig klassifizieren.

Allerdings erlauben es diese Studien nicht, generelle Rückschlüsse auf die Überlegenheit der CBCA- oder RÜ-Kriterien zu ziehen. Vielmehr ist davon auszugehen, dass die Kriterien in Abhängigkeit vom verwendeten Stimulusmaterial unterschiedlich valide sind. Bereits Sporer (1997b) wies darauf hin, dass sich bei der Analyse von sexuellen Missbrauchsfällen durchaus eine Überlegenheit der forensischen gegenüber den RÜ-Kriterien zeigen könnte. Entsprechend diskutierten Strömwall et al. (2004), dass die RÜ-Kriterien möglicherweise einen breiteren Anwendungsbereich aufweisen als die CBCA. Zudem ist erneut darauf hinzuweisen, dass in keiner der dargestellten Untersuchungen alle 19 CBCA-Merkmale beurteilt wurden.

Validität der ARJS

Die Validität der ARJS wurde von Sporer und seiner Arbeitsgruppe in mehreren unabhängigen Studien in Schottland, Deutschland, Australien und Spanien überprüft. Als Stimulusmaterial wurden wahre und erfundene Aussagen von Erwachsenen zu unterschiedlichen Themen verwendet. In Tabelle 3.13 sind die Befunde von vier Studien, in denen trainierte Rater eine vollständige ARJS-

Aussageanalyse vornahmen, zusammenfassend dargestellt. In der Regel ergaben sich erwartungsgemäß höhere Ausprägungen hinsichtlich der ARJS bei wahren als bei erfundenen Aussagen. Am häufigsten erwiesen sich die Skalen zeitliche Informationen, Emotionen und Gefühle, Memorieren und Gedächtnisprozesse sowie Komplikationen und ungewöhnliche Details als valide. Aussagen von Rekruten, die entweder tatsächlich an einer nächtlichen Militärsübung teilgenommen hatten oder dies fälschlicherweise angaben, unterschieden sich zudem erwartungsgemäß hinsichtlich der Skala Fehler und sozial Unerwünschtes (Sporer, 1998). Für diese Skala zeigte sich auch bei Aussagen über persönliche Erlebnisse in der Kindheit oder Jugend ein Effekt geringer Größenordnung (Barnier et al., 2005). In der Untersuchung von Sporer und Burghardt (2004) wiederum ließen sich wahre und erfundene Aussagen über autobiographische Ereignisse auch anhand der Skalen Memorieren und Gedächtnis und Details signifikant voneinander unterscheiden. Die beiden Rater wurden in den Untersuchungen von Sporer (1998) nur kurz und von Sporer und Burghardt (2004) im Rahmen eines 3-wöchigen Trainings auf die ARJS-Aussageanalyse vorbereitet. In der Untersuchung von Sporer und Walther (2006) wurde das Stimulusmaterial von denselben Ratern beurteilt, die bei Sporer und Burghardt die ARJS-Aussageanalyse vorgenommen hatten. Dadurch verfügten die beiden Rater über umfangreiche praktische Vorerfahrungen in der ARJS-Aussageanalyse. Zudem wurden sie nach jeder Beurteilung über den objektiven Wahrheitsstatus der entsprechenden Aussage informiert. Möglicherweise trug diese Rückmeldung dazu bei, dass sich in dieser Studie besonders starke Effekte zeigten. Als Stimulusmaterial verwendeten Sporer und Walther wahre und erfundene Aussagen von Erwachsenen über ihre praktische Führerscheinprüfung. Alle Personen gaben zunächst einen freien Bericht ab und wurden anschließend dazu befragt. Für die freien Berichte differenzierten sieben, für die Interviews neun der 13 ARJS-Skalen signifikant zwischen wahren und erfundenen Aussagen.

Neben den varianzanalytischen Auswertungen wurden in allen Validierungsstudien Diskriminanzanalysen mit den 13 ARJS-Skalen als Prädiktorvariablen und dem objektiven Wahrheitsstatus als Kriteriumsvariable durchgeführt. Die Befunde verweisen auf eine Klassifikationsgüte der ARJS von 64% bis 79%. Teilweise ließen sich wahre (Sporer, 1998; Sporer & Walther, 2006, Interviews), teilweise erfundene Aussagen besser klassifizieren (Barnier et al., 2005; Sporer & Burghardt, 2004; Sporer & Walther, 2006, Freie Berichte).

Die bislang vorliegenden Studien berichteten positive Befunde zur Validität der ARJS. Allerdings sind weitere Untersuchungen erforderlich, um diese Befunde abzusichern. Bisher wurden erst wenige Rater für die ARJS-Aussageanalyse eingesetzt. Daher erscheint es sinnvoll weitere Rater zu schulen, um die Validität der ARJS erneut zu überprüfen.

Moderierende Effekte der Gelegenheit zur Vorbereitung

In der rechtspsychologischen Praxis ist davon auszugehen, dass Personen ihre Aussagen vorbereiten. Daher wurde in mehreren Studien untersucht, ob sich die Gelegenheit zur Vorbereitung auf die Validität inhaltlicher Aussagemerkmale auswirkt.

Alonso-Quecuty (1992) überprüfte den Einfluss der Vorbereitungszeit auf die Validität von drei RÜ-Kriterien. Den Versuchspersonen wurde ein Film gezeigt, den sie sowohl wahrheitsgetreu als auch verfälscht wiedergaben. Der einen Hälfte der Versuchspersonen wurde dabei 10 Minuten Vorbereitungszeit eingeräumt, während die anderen unmittelbar nach der Instruktion aussagten. Wahre Aussagen beinhalteten nur dann mehr sensorische und kontextuelle Informationen als verfälschte, wenn die Probanden keine Gelegenheit hatten sich auf ihre Aussage vorzubereiten.

In anderen Untersuchungen zeigte sich allerdings nicht, dass die Validität inhaltlicher Aussagemerkmale durch die Gelegenheit zur Vorbereitung reduziert wird. Sporer und Küpper (1995) überprüften die Validität von acht RÜ-Kriterien. Die

Probanden schrieben in ausbalancierter Reihenfolge wahre und erfundene Geschichten über persönlich bedeutsame Ereignisse nieder. Dabei erfolgte die erste Niederschrift unmittelbar nach der Instruktion, die zweite eine Woche später. Auf multivariater Ebene interagierte diese Variation der Gelegenheit zur Vorbereitung nicht mit dem Wahrheitsstatus. Allerdings zeigten die univariaten Analysen signifikante Interaktionseffekte für die beiden RÜ-Merkmale Klarheit und räumliche Informationen. Nach einer Woche wurden entgegen der Erwartung deutlichere Mittelwertsunterschiede zwischen wahren und erfundenen Berichten festgestellt, als wenn diese unmittelbar niedergeschrieben wurden.

Ähnliche Befunde wurden von Sporer (1997a) berichtet, der sowohl 13 CBCA- als auch acht RÜ-Kriterien untersuchte. Als Stimulusmaterial wurden mündliche Aussagen zu persönlich bedeutsamen Lebensereignissen verwendet, die später transkribiert wurden. Erneut formulierten alle Probanden sowohl wahre als auch erfundene Aussagen in ausbalancierter Reihenfolge. Die erste Aussage erfolgte unmittelbar, also ohne Vorbereitung. Durch die erste Aussage kam es jedoch zwangsläufig zu einer kurzen zeitlichen Verzögerung für die zweite Aussage. Diese Verzögerung wurde als Gelegenheit zur Vorbereitung aufgefasst. Es ergab sich eine signifikante Wechselwirkung zwischen der Gelegenheit zur Vorbereitung und dem Wahrheitsstatus auf die Aussagequalität. Nach der kurzen zeitlichen Verzögerung zeigten sich deutlichere Mittelwertsunterschiede zwischen wahren und erfundenen Aussagen, als wenn diese unmittelbar nach der Instruktion erfolgten. Dies galt für die vier CBCA-Merkmale Detailreichtum, Raum-zeitliche Verknüpfungen, Schilderung psychischer Vorgänge des Interaktionspartners und nebensächliche Details. Doch auch die beiden RÜ-Merkmale Emotionen und zeitliche Informationen unterlagen diesem Interaktionseffekt.

Zusammenfassend weist die Untersuchung von Alonso-Quecuty (1992) darauf hin, dass die Validität inhaltlicher Aussagemerkmale gemindert wird, während sie sich nach den Befunden von Sporer und Kollegen (Sporer, 1997a;

Sporer & Küpper, 1995) verbessert. Die widersprüchlichen Forschungsbefunde sind möglicherweise auf Unterschiede zwischen den Untersuchungen zurückzuführen (vgl. Masip et al., 2005). So manipulierten Alonso-Quecuty (1992) die Gelegenheit zur Vorbereitung im Rahmen eines Between-Subjects-Designs, während Sporer und Küpper (1995) sowie Sporer (1997a) ein Within-Subjects-Design verwendeten. Zudem variierte die Dauer der Vorbereitung zwischen den drei Studien von einer kurzen Verzögerung (Sporer, 1997a), über 10 Minuten (Alonso-Quecuty, 1992), bis hin zu einer Woche (Sporer & Küpper, 1995).

Zur Frage, ob die ARJS Aussageanalyse auch auf vorbereitete Aussagen anwendbar ist, liegen zwei Studien vor. Sporer (1998) instruierte Probanden entweder einen Abend oder unmittelbar vorher ihre Aussage abzugeben. Für fünf der 13 ARJS-Skalen ergaben sich signifikante Interaktionseffekte zwischen der Vorbereitungszeit und dem Wahrheitsstatus. Die meisten Skalen differenzierten nur dann erwartungsgemäß zwischen wahren und erfundenen Aussagen, wenn den Personen keine Gelegenheit gegeben wurde sich vorzubereiten. Dies zeigte sich für die Skalen Details, zeitliche Details, Gedanken, Memorieren und Gedächtnis, Fehler und sozial Unerwünschtes sowie persönliche Signifikanz. Lediglich für die Skala Komplikationen und ungewöhnliche Details wurden sowohl bei unvorbereiteten als auch bei vorbereiteten Aussagen erwartungsgemäße Mittelwertsunterschiede in Abhängigkeit vom Wahrheitsstatus nachgewiesen. Im Gegensatz dazu berichteten Sporer und Burghardt (2004) zwar höhere ARJS-Beurteilungen nach 15-minütiger als nach 2-minütiger Vorbereitung der Aussage, die Vorbereitungszeit interagierte jedoch nicht mit dem objektiven Wahrheitsstatus. Die Gelegenheit zur Vorbereitung reduzierte demnach nicht die Fähigkeit der ARJS zwischen wahren und erfundenen Aussagen zu unterscheiden.

Die Forschungsbefunde zum moderierenden Einfluss der Gelegenheit zur Vorbereitung auf die Validität der ARJS sind demnach ebenfalls widersprüchlich. Ob die ARJS unabhängig vom Ausmaß der Vorbereitung eine valide Einschätzung des Wahrheitsstatus erlauben, bleibt eine offene Forschungsfrage.

Moderierende Effekte der Valenz der geschilderten Ereignisse

Steller (1989) argumentierte, dass für die experimentelle Überprüfung der CBCA spezielle Anforderungen an das Stimulusmaterial zu stellen sind. Es sollten Aussagen zu Ereignissen verwendet werden, bei denen die aussagende Person direkt beteiligt war, die eine negative Valenz aufwiesen und mit einem Gefühl von Kontrollverlust einhergingen. Diesen Forderungen entsprechend wurde die Validität inhaltlicher Aussagemerkmale meist anhand von Aussagen zu negativen Ereignissen untersucht (z.B. Landry & Brigham, 1992). Dies erschien notwendig, um eine Generalisierung von Forschungsbefunden auf das praktische Anwendungsfeld des inkriminierten sexuellen Missbrauchs zu erlauben. Allerdings bleibt offen, ob die inhaltliche Aussageanalyse auch auf positive Ereignisse anwendbar ist. Die ARJS beanspruchen einen breiten Geltungsbereich, der über Sexualdelikte hinausgeht (Sporer, 2004). Daher stellt sich die Frage, ob inhaltliche Aussagemerkmale auch dazu geeignet sind, zwischen wahren und erfundenen Berichte über positive Ereignisse zu unterscheiden.

Verschiedene Untersuchungen zum RÜ-Ansatz haben Unterschiede in der Selbstbeschreibung von Erinnerungsqualitäten in Abhängigkeit von der Valenz des Ereignisses festgestellt. Wiederholt wurden positive Ereignisse hinsichtlich verschiedener RÜ-Merkmale höher beurteilt als negative Ereignisse (D'Argembeau, Comblain & van der Linden, 2003; Destun & Kuiper, 1999; Larsen, 1998; Schaefer & Philippot, 2005). Während sich positive Erinnerungen deutlich von neutralen unterschieden, wurden kaum Unterschiede zwischen negativen und neutralen Erinnerungen nachgewiesen (D'Argembeau et al., 2003; Schaefer & Philippot, 2005). Die Autoren führten erklärend aus, dass Versuchspersonen es im Rahmen von Laborstudien möglicherweise vermeiden, sich detailliert an negative Ereignisse zu erinnern. Diese Vermeidungstendenz ließe sich lediglich dann überwinden, wenn spezifische Abrufhinweise vorgegeben würden. Entsprechend argumentierte auch Christianson (1992), dass im Rahmen von

freien Berichten negative Ereignisse weniger gut zugänglich sind als neutrale. Werden jedoch verschiedene Arten von Abrufhinweisen vorgegeben, wie es beispielsweise im Rahmen von Interviews geschieht, scheinen Unterschiede im Abruf emotionaler und neutraler Ereignisse zu verschwinden (Christianson, 1992).

Destun und Kuiper (1999) untersuchten nicht nur Erinnerungen an erlebnisbasierte Ereignisse, sondern instruierten die Probanden auch sich Ereignisse mit positiver oder negativer Valenz vorzustellen. Für die Selbstbeurteilung von Emotionen ergab sich eine Wechselwirkung zwischen der Valenz und der Erfahrungsgrundlage des Ereignisses. Es zeigten sich deutlichere Unterschiede zwischen erlebnisbasierten und vorgestellten Ereignissen, wenn diese als unangenehm empfunden wurden, als wenn sie angenehm waren. Zudem schienen vorgestellte Ereignisse vor allem Gefühle zu beinhalten, die mit der Valenz des Ereignisses kongruent waren. Im Gegensatz dazu umfassten erlebnisbasierte Ereignisse auch Emotionen, die im Widerspruch zu der vorherrschenden Valenz standen. So wurden für vorgestellte angenehme Ereignisse mehr positive Emotionen berichtet, als für erlebnisbasierte angenehme Ereignisse. Zudem wiesen erlebnisbasierte Ereignisse im Gegensatz zu vorgestellten auch dann ein gewisses Maß an positiven Gefühlen auf, wenn sie unangenehm waren. Destun und Kuiper führten interpretierend aus, dass es aufgrund dieser Wechselwirkung schwierig sein könnte zwischen vorgestellten angenehmen und erlebnisbasierten unangenehmen Ereignissen zu differenzieren.

Larsen (1998) wiederum fand Wechselwirkungseffekte zwischen der Erlebnisgrundlage und der Valenz auf die Selbstbeurteilung der Erinnerungsqualität für das RÜ-Merkmal Lebendigkeit. Positive Ereignisse wurden als lebendiger beschrieben als negative. Auch zeigte sich, dass erlebnisbasierte Ereignisse im Vergleich zu erfundenen als lebendiger wahrgenommen wurden**.** Die Variation der Valenz erfolgte unter anderem dadurch, dass den Probanden spezifische Gefühlszustände vorgegeben wurden, die mit dem Ereignis assoziiert

sein sollten. Dabei zeigten sich für negative Ereignisse erneut deutlichere Unterschiede in der Lebendigkeit erlebnisbasierter und erfundener Ereignisse als für positive. Aufgrund der geringen Effekte schlussfolgerte Larsen jedoch, dass das RÜ-Merkmal Lebendigkeit möglicherweise weniger gut zur Unterscheidung wahrer und vorgestellter Gedächtnisinhalte geeignet ist als andere Merkmale. Eine höhere Erinnerungsqualität für positive im Vergleich zu negativen Ereignissen steht im Einklang mit dem in der Gedächtnispsychologie diskutierten Polyanna-Prinzip bzw. Positivitätsbias (vgl. Matlin, 2004; Pohl, 2007). Dabei wird davon ausgegangen, dass aufgrund von stimmungsregulierenden Prozessen (z.B. Kennedey, Mather & Carstensen, 2006) positive autobiographische Ereignisse besser erinnert werden als negative. Nach dem Polyanna-Prinzip wäre zudem zu erwarten, dass Probanden bei freier Themenwahl eher positive als negative Ereignisse berichten. Im Gegensatz dazu fanden Sporer und Sharman (2006) sowie Sporer und Küpper (1995, 2004), dass ohne konkrete Themenvorgabe gleich häufig negative und positive Lebensereignisse geschildert wurden.

Sporer und Sharman (2006) instruierten Studierende aufregende oder besondere Lebensereignisse entweder wahrheitsgemäß darzustellen oder frei zu erfinden. Zunächst stellten die Studierenden die frei gewählten Ereignisse schriftlich dar. Danach bewerteten sie ihre eigenen Darstellungen anhand von mehreren RÜ-Merkmalen, die nachträglich sieben RÜ-Skalen zugeordnet wurden. Für vier der sieben RÜ-Skalen waren signifikante Haupteffekte der anhand von Selbstbeschreibungen erfassten Valenz der Ereignisse festzustellen. Zudem zeigte sich eine Wechselwirkung zwischen der Valenz und dem Wahrheitsstatus der geschilderten Ereignisse auf die Selbstbeurteilung von räumlichen Informationen. Es ergaben sich nur für negative Ereignisse erwartungsgemäße Mittelwertsunterschiede zwischen wahren und erfundenen Ereignissen. Zur Entdeckung von Täuschung ist jedoch die Frage relevant, ob sich die selbstberichteten Unterschiede in den Erinnerungsqualitäten auch in den

Darstellungen der Ereignisse abbilden. Beispielsweise fanden Bohanek, Fivush und Walker (2005) keine Zusammenhänge zwischen den Selbstbeurteilungen von persönlichen Lebensereignissen und der Art und Weise wie diese dargestellt wurden. Allerdings bezogen sich die Selbstbeurteilungen auf Erinnerungsqualitäten (z.B. Lebendigkeit), während die Art der Darstellung vor allem anhand linguistischer Indikatoren (z.B. grammatikalische und semantische Struktur) erfasst wurde. Relevanter erscheint es jedoch zu überprüfen, ob Selbst- und Fremdbeurteilungen hinsichtlich derselben inhaltlichen Merkmale korrespondieren. Daher untersuchten Sporer und Sharman, ob sich die selbstberichteten Unterschiede in Abhängigkeit von der Valenz der Ereignisse auch anhand von Fremdbeurteilungen erfassen lassen. Die Autoren ließen jeden Versuchsteilnehmer die schriftliche Darstellung eines anderen Teilnehmers anhand derselben RÜ-Merkmale einschätzen, die bereits für die Selbstbeurteilungen verwendet wurden. Für die so gewonnenen Fremdbeurteilungen ergab sich zwar ebenfalls ein signifikanter Haupteffekt der Valenz auf die RÜ-Skala Gedächtnisqualität und Memorieren, Wechselwirkungen zwischen der Valenz und dem Wahrheitsstatus waren jedoch nicht festzustellen.

Auch Barnier et al. (2005) überprüften, ob sich die Befunde von Selbstbeurteilungsstudien auf die interpersonelle Ebene übertragen lassen. Zur experimentellen Manipulation der Erfahrungsgrundlage ließen sie jeden Probanden ein erlebnisbasiertes und zwei erfundene (ein vorgestelltes und ein gezielt vorgetäuschtes) Ereignisse aufschreiben. Zudem wurden die Probanden instruiert entweder positive oder negative Ereignisse zu schildern. Sämtliche Darstellungen wurden dann durch einen Rater anhand der ARJS und fünf RÜ-Kriterien analysiert.[2] Diese Fremdbeurteilung der Aussagequalität unterschied

---

[2] Die Validität der ARJS- und RÜ-Kriterien vergleichend zu interpretieren ist leider nicht möglich, da die Aussageanalysen durch jeweils einen Rater vorgenommen wurden (vgl. Wells & Windschitl, 1999). Ein Teil des Stimulusmaterials wurde zwar

sich nicht nur in Abhängigkeit von der Erfahrungsgrundlage, sondern auch von der Valenz des dargestellten Ereignisses. Allerdings wurde im Gegensatz zu den Befunden von Selbstbeurteilungsstudien die Aussagequalität der negativen Ereignisse durch die Rater meist höher eingeschätzt als die der positiven. Dies galt sowohl für die Beurteilung anhand der ARJS- als auch der RÜ-Kriterien. Eine Wechselwirkung zwischen der Erfahrungsgrundlage und der Valenz war nicht nachweisbar. Demnach waren die Inhaltskriterien sowohl für positive als auch für negative Ereignisse valide.

Die Befunde von Barnier et al. (2005) sowie von Sporer und Sharman (2006) legen nahe, dass auch Fremdbeurteilungen der Aussagequalität durch die Valenz des Ereignisses beeinflusst werden. Dennoch bildeten die Fremdbeurteilungen der inhaltlichen Aussagemerkmale unabhängig von der Valenz der Ereignisse erwartungsgemäße Unterschiede zwischen wahren und erfundenen Aussagen ab. Allerdings stellten die Probanden in beiden Untersuchungen die relevanten Ereignisse schriftlich dar. Es ist jedoch denkbar, dass schriftliche Darstellungen von Erinnerungen genauer überarbeitet werden als mündliche Aussagen (vgl. Johnson, 1988) und dies Unterschiede in der inhaltlichen Qualität bewirkt. Demnach erscheinen weitere Untersuchungen erforderlich, um die Befunde der bisherigen Studien abzusichern.

Moderierende Effekte der Befragungsform

Auch die Interaktion mit einem Interviewer sowie die Befragungsform könnten sich auf die Validität inhaltlicher Aussagemerkmale auswirken. Um Einflüsse der Befragung auszuschließen, wäre es am sichersten die Inhaltsanalyse auf frei formulierte Aussagen zu beschränken. So setzt beispielsweise eine Beurteilung des CBCA-Kriteriums der unstrukturierten

---

von einem weiteren Rater analysiert, doch für einige Skalen wurden nur geringe Inter-Rater-Reliabilitäten nachgewiesen.

Darstellung voraus, dass die Aussagen nicht zu sehr durch eine direkte Befragung strukturiert wurden (Steller & Köhnken, 1989). Aus praktischen Überlegungen heraus läßt sich jedoch nicht auf eine Befragung verzichten.

So erfordert eine inhaltliche Aussageanalyse, dass umfangreiche Aussagen vorliegen. Daher erscheint es notwendig, insbesondere bei der Befragung von Kindern, Abrufhinweise vorzugeben (Steller & Boychuk, 1992). In der Regel wird jedoch davon ausgegangen, dass die inhaltliche Aussageanalyse auch auf Aussagen anwendbar ist, die im Rahmen nicht-suggestiver Befragungen erhoben wurden (z.B. Steller & Köhnken, 1989). Strittiger ist hingegen die Frage, ob das sogenannte Kognitive Interview nach Geiselman, Fisher, MacKinnon und Holland (1986; vgl. auch Fisher & Geiselman, 1992) als spezifische Interviewtechnik die Validität der inhaltlichen Aussageanalyse beeinträchtigt.

Steller und Wellershaus (1996) führten aus, dass einzelne Techniken des Kognitiven Interviews die anhand von CBCA-Merkmalen erfasste Qualität kindlicher Aussagen erhöhen. Zudem zeigte sich, dass Laien erfundene Aussagen häufiger fälschlicherweise als glaubhaft beurteilten, wenn sie durch Kognitive Interviews im Vergleich zu Standardinterviews erhoben wurden. Die Autoren führten diesen Befund auf die erhöhte Aussagequalität zurück und warnten daher davor, die CBCA auf Aussagen anzuwenden, die durch Kognitive Interviews erhoben wurden. Colwell et al. (2002) sowie Köhnken, Schimossek, Aschermann und Höfer (1995) fanden jedoch keine Wechselwirkungen zwischen der Art des Interviews und dem Wahrheitsstatus auf die untersuchten CBCA-Kriterien. Sie schlussfolgerten daher, dass die Validität der inhaltlichen Aussagemerkmale nicht durch das Kognitive Interview gefährdet wird.

Auch die Validität von RÜ-Merkmalen wurde in Abhängigkeit von der Befragungstechnik untersucht (Hernandez-Fernaud & Alonso-Quecuty, 1997; Larsson & Granhag, 2005). Hernandez-Fernaud und Alonso-Quecuty (1997) fanden, dass die beiden RÜ-Merkmale kontextuelle und sensorische Informationen besser zwischen wahren und erfundenen Aussagen unterschieden,

wenn diese über ein Kognitives Interview erhoben wurden im Vergleich zu einem polizeilichen Standardinterview. Larsson und Granhag (2005) verglichen Aussagen, die über ein Kognitives und über ein Strukturiertes Interview erhoben wurden anhand von sieben RÜ-Kriterien. Im Rahmen des Kognitiven Interviews wurden sowohl erlebnisbasierte als auch erfundene Ereignisse berichtet, während alle über das Standardinterview erhobenen Aussagen erfunden waren. Aus dem Vergleich dieser drei Untersuchungsbedingungen schlussfolgerten die Autoren, dass das Kognitive Interview die Validität von RÜ-Kriterien reduzieren könnte, ohne sie jedoch aufzuheben. Allerdings fehlte eine Vergleichsbedingung, in der erlebnisbasierte Aussagen über ein Standardinterview erhoben wurden, um diese Interpretation der Untersuchungsbefunde empirisch abzusichern.

Sporer und Walther (2006) wiederum verglichen die Validität der ARJS für Aussagen, die frei berichtet oder im Rahmen eines standardisierten Interviews geschildert wurden. Es zeigten sich jedoch kaum Unterschiede in der Validität der inhaltlichen Aussagemerkmale. Die Interviews erfolgten entweder anhand von allgemeinen Fragen, z.B. nach dem genauen Zeitpunkt und Ort des Geschehens, oder anhand von Fragen, die direkt aus den ARJS abgeleitet wurden. Diese Variation der Befragungstechnik wirkte sich zwar nicht auf die Validität der ARJS insgesamt aus, für zwei Skalen ergaben sich jedoch Wechselwirkungseffekte. Wahre und erfundene Aussagen ließen sich hinsichtlich der Skala außergewöhnliche Details besser voneinander unterscheiden, wenn allgemeine Fragen gestellt wurden. Hingegen erwies sich die Skala Fehler und sozial Unerwünschtes bei einer Befragung anhand der ARJS als valider.

Insgesamt legen die Befunde nahe, dass spezifische Befragungstechniken, z.B. des Kognitiven Interviews, die Aussagequalität beeinflussen können. Dass die Validität inhaltlicher Aussagemerkmale dadurch vollständig aufgehoben wird, ist aufgrund der vorliegenden Untersuchungsbefunde jedoch nicht zu befürchten.

Ziele und Hypothesen

Die vorliegende Untersuchung zielte darauf ab, die Reliabilität und Validität der ARJS erneut zu überprüfen. Zudem sollte der Einfluss potenzieller Moderatorvariablen auf die Fremdbeurteilung von Aussagen anhand der ARJS-Merkmale untersucht werden.

In Anlehnung an bisherige Forschungsbefunde ist zu erwarten, dass ARJS-Beurteilungen eine höhere Aussagequalität bei wahren im Vergleich zu erfundenen Aussagen abbilden. Zudem wird angenommen, dass sich die Aussagequalität durch Vorbereitung verbessern lässt. Entsprechend sollten vorbereitete Aussagen höhere ARJS-Beurteilungen erhalten als unvorbereitete. Des Weiteren bleibt zu klären, ob die Validität der ARJS durch die Gelegenheit zur Vorbereitung vermindert wird (vgl. Sporer, 1998) oder diesem Einfluss gegenüber robust ist (vgl. Sporer & Burghardt, 2004).

Studien zum RÜ-Ansatz berichteten höhere Selbstbeurteilungen der Aussagequalität bei positiven als bei negativen Ereignissen (D'Argembeau et al., 2003; Destun & Kuiper, 1999; Larsen, 1998; Schaefer & Philippot, 2005). Für Fremdbeurteilungen ist allerdings gemäß der Befunde von Barnier et al. (2005) Gegenteiliges zu erwarten. Daher wird postuliert, dass negative Ereignisse höhere ARJS-Beurteilungen aufweisen als positive. Ebenso soll überprüft werden, ob die ARJS unabhängig von der Valenz des Ereignisses dazu geeignet sind wahre und erfundene Aussagen zu unterscheiden (vgl. Barnier et al., 2005). Dies würde die Zielsetzung eines breiten Geltungsbereich der ARJS unterstützen.

Schließlich liegen Untersuchungsbefunde vor, die zeigen, dass die Validität inhaltlicher Aussagemerkmale durch die spezifischen Befragungstechniken des Kognitiven Interviews nicht reduziert wird (Colwell et al., 2002; Köhnken et al., 1995; Hernandez-Fernaud & Alonso-Quecuty, 1997). Auch wenn Steller und Wellershaus (1996) widersprüchliche Befunde für CBCA-Merkmale berichteten, wurden für die Validität der ARJS bislang keine Unterschiede zwischen freien Berichten und Interviews festgestellt (Sporer & Walther, 2006). Daher wird

angenommen, dass die Validität der ARJS nicht auf die Analyse freier Berichte beschränkt ist, sondern auch für Aussagen nachzuweisen ist, die im Rahmen von Interviews gewonnen wurden.

Methode

Stimulusmaterial

Als zu beurteilendes Stimulusmaterial wurden Aussagen von 176 Frauen aus der Studie von Sporer und Burghardt (2004) verwendet. Der Erhebung des Stimulusmaterials lag ein 2 x 2 x 2 (x 2) Design zugrunde mit den Between-Subjects-Faktoren Wahrheitsstatus (erfunden versus wahr), Vorbereitungszeit (2 versus 15 Minuten) und Motivation (gering versus mittel) sowie dem Within-Subjects-Faktor Aussageform (freier Bericht versus Interview). Die Probandinnen wurden instruiert entweder ein wahres oder erfundenes Ereignis zu berichten. Dazu wählten sie eines von 10 vorgegebenen Themen aus. Diese umfassten emotional relevante Ereignisse sowohl mit positiver Valenz, wie die Geburt eines Kindes oder die eigene Hochzeit, als auch mit negativer Valenz, beispielsweise einen Krankenhausaufenthalt oder den Tod einer nahestehenden Person. Jede Probandin gab zunächst einen freien Bericht über das Thema ab. Dabei wurde die Gelegenheit zur Vorbereitung experimentell manipuliert. Die Hälfte der Probandinnen hatte 2 Minuten, die andere Hälfte 15 Minuten Zeit, sich auf ihre Aussage vorzubereiten. Zudem wurde der Hälfte der Probandinnen ein finanzieller Anreiz für eine überzeugende Aussage in Aussicht gestellt. Aufgrund eines Manipulation-Checks ist jedoch anzunehmen, dass das Ausmaß an Motivation dadurch nicht erfolgreich variiert wurde (vgl. Sporer & Zander, 2001). Dieser Faktor wurde daher in der vorliegenden Untersuchung nicht weiter berücksichtigt. Eine Woche später wiederholten die Probandinnen ihre Aussage während eines Interviews. Dabei wurden ihnen Anschlussfragen gestellt.

Pilotstudie

Da die Versuchspersonen semi-strukturiert befragt wurden, erschien es sinnvoll, nachträglich zu analysieren welche Fragen gestellt wurden. Um die Fragen der Interviewerin genauer zu beschreiben, wurden sie von zwei Raterinnen kategorisiert. Diese Raterinnen waren ansonsten nicht an der Studie beteiligt, hinsichtlich des Wahrheitsstatus der einzelnen Transkripte blind und wurden über weitere experimentelle Manipulationen nicht informiert. Jede von ihnen beurteilte die Hälfte der Interviews mit Hilfe einer Software zur qualitativen Datenanalyse (MAXQDA, www.maxqda.de). Diese ermöglichte es, die Fragen in ihrem Kontext zu bewerten. Von besonderem Interesse erschien es zu überprüfen, ob spezifische Fragen der Interviewerin die Ausprägung einzelner ARJS-Skalen beeinflusst haben könnten.

Daher konzipierte die Verfasserin zunächst Beurteilungskategorien auf der Grundlage einzelner ARJS-Skalen. Für einzelne Skalen erschienen spezifische Nachfragen naheliegend. Beispielsweise könnten sich recht gebräuchliche offene Fragen, die durch wo, wann oder wer eingeleitet werden, auf die Skalen räumliche Details, zeitliche Details, nonverbale und verbale Interaktionen auswirken. Hingegen schien es unwahrscheinlich, dass die Interviewerin gezielt nachfragen würde, ob die Personen sich klar an das geschilderte Ereignis erinnern würde. Diese Überlegungen resultierten zunächst in neun Beurteilungskategorien, die sich eng an den ARJS-Skalen orientierten. Anhand dieser Kategorien beurteilten die beiden Raterinnen zunächst acht Aussagen. Diese Aussagen stammten aus derselben Untersuchung wie das eigentliche Stimulusmaterial (Sporer & Burghardt, 2004), wurden jedoch in der vorliegenden Untersuchung nicht verwendet. Danach erfolgte eine intensive Diskussion zwischen den beiden Raterinnen und der Verfasserin der vorliegenden Arbeit. Dabei wurden die ursprünglichen Kategorien teilweise aufgegeben, ergänzt und modifiziert. Schließlich resultierten zwölf Kategorien, die einen Großteil der gestellten Fragen abdecken sollten. Diese erhielten die Bezeichnungen Klarheit und Lebendigkeit,

räumliche Details, zeitliche Details, Emotionen und Gefühle, Memorieren und Gedächtnis, Interaktionen, Präzision von Personen und Tieren, von Abläufen, von Objekten, von Namen, allgemeine Ergänzungen und Prozedurales.

Ein Teil der Interviews (32 von 176, 18%) wurde von beiden Raterinnen unabhängig beurteilt. Dadurch war es möglich, die Inter-Rater-Reliabilität der Kategorisierungen abzuschätzen. Als zufallskorrigiertes Maß der Beurteilerübereinstimmung wurde Cohen's kappa berechnet. Allerdings waren explizite Aufforderungen, Gefühle zu schildern, nicht in dem Teil des Stimulusmaterials vorzufinden, der von beiden Raterinnen bearbeitet wurde, und für drei weitere Kategorien lagen asymmetrische Häufigkeitsverteilungen vor. Daher wurde zusätzlich die prozentuale Häufigkeit absoluter Übereinstimmungen (PÜ) ermittelt.

Die Interviewerin stellte durchschnittlich sieben Fragen pro Interview (M = 7.17, SD = 2.21, Min = 3, Max = 15; Anzahl der Interviewer-Fragen insgesamt: N(I) = 1262). In der Regel wurde die offene Frage gestellt, ob die Probandin ihrer Aussage noch etwas hinzuzufügen habe (n(I) = 257, Anteil an allen Fragen: 20.36%; PÜ = 81.3). Zudem wurden die Probandinnen häufig dazu aufgefordert, zeitliche (z.B. „Wie lange hat der Geburtsvorgang gedauert?“, n(I) = 311, 24.64%; Cohen's kappa = .90, PÜ = 93.8) und räumliche Details („Von wo bis wo sollte damals die Kanutour gehen?“, n(I) = 197, 15.61%; Cohen's kappa = 1, PÜ = 100) zu ergänzen. Des Weiteren wurden sie nach potenziellen Interaktionspartnern gefragt (z.B. „Wie viele Personen waren da?“, n(I) = 131, 10.38%; Cohen's kappa = .94, PÜ = 96.9), nach konkreten Namen, (z.B. „Wie war der Name des Tierarztes, der Rudi dann behandelt hat?“, n(I) = 57, 4.52%; Cohen's kappa = 1, PÜ = 100) oder nach Personenbeschreibungen (z.B. „Wie sahen irgendwelche der bedeutenden Gäste aus?“, n(I) = 39, 3.09%; Cohen's kappa = .89, PÜ = 96.9). Auch wurden sie gebeten, Objekte (z.B. „Können Sie sich denn noch erinnern, wie genau das Fahrzeug aussah?“, n(I) = 59, 4.67, PÜ = 93.7) oder Abläufe (z.B. „Können Sie den Unfallhergang nochmal genauer beschreiben?“, n(I) = 66, 5.23%;

PÜ = 84.4) zu präzisieren. Weitere Fragen zielten darauf ab zu erfahren, was vor oder nach dem geschilderten Ereignis statt gefunden hat (z.B. „Was ist im Anschluss der Beerdigung gewesen?“, n(I) = 49, 3.88%; Cohen’s kappa = .78, PÜ = 90.6). Zuweilen forderte die Interviewerin die Probanden dazu auf, sich in konkrete Situationen hineinzuversetzen (z.B. „Versuchen Sie sich das bildhaft vorzustellen“,n(I) = 25, 1.98%; Cohen’s kappa = 1, PÜ = 100). Nur selten wurden die Probandinnen explizit nach ihren Gefühlen in der geschilderten Situation gefragt (n(I) = 7, 0.55%; PÜ = 100). Die übrigen Äußerungen der Interviewerin bezogen sich auf das experimentelle Prozedere oder dienten der Herstellung einer angenehmen Gesprächsatmosphäre (z.B. durch das Paraphrasieren von vorangegangenen Äußerungen der Probandin, n(I) = 64, 5.07%; Cohen’s kappa = .45, PÜ = 68.8). Diese Kategorie war die einzige, für die keine sehr gute Inter-Rater-Reliabilität nachgewiesen werden konnte. Da die Kategorisierungen jedoch lediglich der Beschreibung des Stimulusmaterials dienten, ist die Beurteilerübereinstimmung durchaus noch zu tolerieren (vgl. Wirtz & Caspar, 2002). Zudem wurden für alle anderen Kategorien Werte erzielt, die als äußerst gut einzustufen sind, alle Cohen’s kappas ≥ .75, alle PÜs ≥ 81.3.

Für die vorliegende Untersuchung wurden die Transkripte von 176 freien Berichten und 176 Interviews anhand der ARJS beurteilt. Das verwendete Stimulusmaterial war insgesamt sehr umfangreich und die Anzahl der Wörter variierte stark. Die freien Berichte umfassten im Durchschnitt 767 Wörter (SD = 449, Min = 179, Max = 2567), die Interviews exklusive der Fragen 1011 Wörter (SD = 535, Min = 196, Max = 3073).

Beurteilertraining

Vier Raterinnen wurden intensiv im Umgang mit den ARJS trainiert. Alle studierten Psychologie im Hauptstudium und hatten bereits im Rahmen eines Seminars relevante Vorkenntnisse zum Thema Täuschung und Entdeckung von Täuschung erworben. Das Training untergliederte sich in ein vorbereitendes

Literaturstudium, eine theoretische Einführung und eine Übungsphase. Zur unmittelbaren Vorbereitung auf das Training bearbeiteten die Raterinnen eigenständig Literatur zu sozialpsychologischen Aspekten von Täuschung (Fiedler, 1989a), zur CBCA (Steller & Köhnken, 1989; Vrij, 2005) sowie zum RÜ-Ansatz (Johnson & Raye, 1981; Masip et al., 2005). Im Rahmen einer fünftägigen theoretischen Einführung durch die Verfasserin wurden relevante Forschungsbefunde erläutert und das ARJS-Manual (Sporer, 1996/1998/2004) besprochen. Das Manual umfasst allgemeine Instruktionen zum Einsatz der ARJS, Definitionen der ARJS-Skalen und Items sowie Regeln zur Ableitung eines abschließenden Gesamturteils zur Glaubhaftigkeit. Sämtliche Items sind durch Beispiele illustriert und in der Regel wird kurz darauf hingewiesen warum einzelne Merkmale eher bei wahren als bei erfundenen Aussagen erwartet werden. Während der Übungsphase beurteilten die Raterinnen 14 Transkripte anhand der ARJS zunächst individuell und diskutierten anschließend Beurteilungsdiskrepanzen aus. Das Training beanspruchte insgesamt eine Dauer von 50 Stunden.

Durchführung der ARJS-Aussageanalyse

Die Raterinnen erhielten keinerlei Angaben zu dem Anteil wahrer und erfundener Aussagen im Stimulusmaterial und wurden über die weiteren experimentellen Manipulationen nicht informiert. Jede Raterin beurteilte jeweils die Hälfte der freien Berichte und alle Interviews. Um Reihenfolgeeffekte auszubalancieren, bearbeiteten jeweils zwei Raterinnen dieselben Aussagen in umgekehrter Reihenfolge.

Die Aussageanalyse anhand der ARJS erfolgte gemäß den Ausführungen im Manual (Sporer, 1996/1998/2004). Zunächst sind alle 52 Items anhand von Ratingskalen mit Werten von 1 bis 7 zu beurteilen. Niedrige Werte indizieren eine geringe und hohe eine starke Ausprägung des entprechenden Merkmals. Die Endpunkte der Ratingskalen sind entweder qualitativ (z.B. vage--klar) oder

quantitativ (z.B. keine--einige) verankert. Die Ratingskalen sind 7-stufig oder auf drei bis vier Stufen reduziert (1--4--7 oder 1--3--5--7). Dadurch werden Unterschiede in der vermeintlichen Basisrate der einzelnen Merkmale berücksichtigt. Die eingeschränkten Ratingskalen werden für seltene Aussagemerkmale verwendet, so dass deren Auftreten automatisch zu einer relativ hohen Beurteilung führt.

Die Itembeurteilungen werden dann zu 15 Skalenbewertungen zusammengefasst. Die Abweichung von den empirisch ermittelten 13-ARJS-Skalen ergibt sich daraus, dass für die Skalen Komplikationen und ungewöhnliche Details sowie Fehler und Sozial Unerwünschtes gemäß ihrer Bezeichnung jeweils zwei Unterkategorien gebildet werden. Die getrennte Bewertung dieser vier Aspekte unterstützt eine hohe Gewichtung der Merkmale. Die Skalenbewertungen werden wiederum zu drei Skalenblöcken integriert. Sowohl die Skalen als auch die Skalenblöcke werden anhand 3-stufiger Skalen beurteilt. Schließlich werden die Skalenblöcke anhand vorgegebener Gewichtungsrichtlinien zu einem Gesamturteil über die Glaubhaftigkeit der Aussage integriert**.** Am geringsten gewichtet wird der Skalenblock, der die beiden Skalen Realismus und logische Struktur sowie Klarheit und Lebendigkeit umfasst. Ein mittleres Gewicht erhält die Zusammenfassung der Skalen, die sich auf Details, Sinneseindrücke, Emotionen und Gefühle, Gedanken, Memorieren und Gedächtnis sowie Interaktionen beziehen. Das stärkste Gewicht erhält der Block, der die Skalenbewertungen für Komplikationen, ungewöhnliche Details, Fehler und sozial Unerwünschtes integriert.

Das Glaubhaftigkeitsurteil wird über eine 10-stufige Skala mit den Endpunkten 1 = erfunden/verfälscht und 10 = erlebt/wahr erhoben. Werte von 1 bis 5 sind demnach als Lügenurteile, Werte von 6 bis 10 als Wahrheitsurteile aufzufassen. Des Weiteren ist die subjektive Sicherheit hinsichtlich des Glaubhaftigkeitsurteils auf einer 5-stufigen Skala mit den Endpunkten 1 = sehr unsicher bis 5 = sehr sicher anzugeben. Die Eignung des Stimulusmaterials für

die inhaltliche Aussageanalyse wird auf einer 10-stufigen Skala beurteilt. Extrem niedrige Werte verweisen darauf, dass eine Aussage überhaupt nicht für die Inhaltsanalyse geeignet ist, während extrem hohe Werte eine sehr gute Eignung indizieren. Im Rahmen der vorliegenden Untersuchung gaben die Raterinnen zusätzlich für jede Aussage an, wie sie die Valenz des geschilderten Ereignisses einschätzten. Dazu wurde eine 7-stufige Skala mit den Endpunkten 1 = negativ und 7 = positiv verwendet, so dass hohe Werte eine positive Erlebnisqualität indizieren.

Ergebnisse

Im Folgenden wird zunächst das Stimulusmaterial hinsichtlich seiner Valenz, seiner Eignung für die inhaltliche Aussageanalyse sowie seines Umfangs beschrieben. Danach werden die Befunde zur Inter-Rater-Reliabilität für die ARJS-Beurteilungen vorgestellt. Im Anschluss daran werden die Befunde zur Validität der ARJS dargestellt. Dazu wurden die Effekte des Wahrheitsstatus (erfunden vs. wahr) und der Vorbereitungszeit (2 vs. 15 Minuten) auf die ARJS-Beurteilungen anhand von 2 x 2 MANOVAs untersucht. Neben dem Wahrheitstatus und der Vorbereitungszeit wurden auch die Effekte der Aussageform (freie Berichte vs. Interviews) anhand einer 2 x 2 (x 2) MANOVA analysiert. Die Ergebnisdarstellung abschließend wird auf die varianzanalytischen Befunde zur Valenz der geschilderten Ereignisse eingegangen.

Wenn varianzanalytisch ein multivariat signifikanter Effekt auf die ARJS-Beurteilungen festzustellen war, wurden zudem univariate Analysen durchgeführt. Um die Stärke der Effekte auf univariater Ebene abzuschätzen, wurden für die F-Werte nach der Formel von Mullen (1989, vgl. auch Rosenthal, 1994) $\text{Direction of Effect} * (F/(F+df_{denom}))^{0.5}$ die Effektstärkemaße r bestimmt. Diese entsprechen den punktbiserialen Korrelationen zwischen dem jeweiligen dichotomen Faktor und der intervallskalierten abhängigen Variable. Cohen (1988) schlug Richtlinien zur Interpretation verschiedener Effektstärkemaße vor. Wendet man diese auf die punktbiseriale Korrelation an, so indiziert ein Wert von $r_{pb} = .37$ einen starken, $r_{pb} =$

.24 einen mittleren und r_{pb} = .10 einen schwachen Effekt (vgl. auch Rice & Harris, 2005). Für univariate Analysen von Wechselwirkungseffekten wurde das Effektstärkemaß f gemäß der Formel von Cohen (1988) $f = (eta^2 /(1-eta^2))^{0.5}$ abgeleitet. Dabei ist f = .10 als geringer, f = .25 als mittlerer und f = .40 als starker Effekt zu interpretieren (vgl. Cohen, 1988; Stelzl, 2005).

Um die Effektstärke für den Messwiederholungsfaktor der Aussageform zu bestimmen, wurde zunächst Cohen's d berechnet (vgl. Dunlap, Cortina, Vaslow & Burke, 1996; Lipsey & Wilson, 2001). Gemäß der Formel $d=(M_1-M_2)/SD_{pooled}$ (Cohen, 1988) wird dabei die Differenz der Mittelwerte für die freien Berichte und die Interviews an deren gepoolter Standardabweichung relativiert. Um die Vergleichbarkeit mit den anderen Untersuchungsbefunden zu erleichtern, wurde Cohen's d danach anhand der Formel von Mullen (1989) in r transformiert ($r=d/(SQRT(d^2+4))$).

Voranalysen zur Valenz, Eignung und Länge der Aussagen

Um das Stimulusmaterial genauer zu beschreiben, schätzten die Raterinnen die Valenz der berichteten Ereignisse und die Eignung der Aussagen für eine inhaltliche Aussageanalyse ein. Zudem wurde für jedes Transkript die Anzahl der Wörter als Indikator der Aussagelänge ausgezählt.

Die 7-stufige Antwortskala zur Einschätzung der Valenz der berichteten Ereignisse wurde voll ausgeschöpft. Nur 12 der 176 (6.8%) freien Berichte erhielten von beiden Raterinnen einen Skalenwert von vier, der auf eine neutrale Valenz verweist. Insgesamt wurden deutlich mehr negative (n = 136 = 77.3% für M < 4) als positive Ereignisse (n = 28 = 15.9% für M > 4) geschildert, M = 3.13, SD = 1.57. Da sich die Interviews auf dieselben Ereignisse bezogen wie die freien Berichte, überrascht es kaum, dass deren Valenz ähnlich eingeschätzt wurde (n = 138 = 78.4% für M < 4, n = 31 = 17.6% für M > 4, n = 7 = 4% für M = 4, M = 3.12, SD = 1.45). Dies spiegelte sich in einer hohen Korrelation zwischen den

Valenzurteilen für die freien Berichte und die Interviews wider, r(174) = .91, p < .001.

Die Aussagen waren für die inhaltliche Analyse unterschiedlich gut geeignet (M = 5.30; SD = 1.80; Min = 1.50; Max = 9.38). Eine 2 x 2 (x 2) Mixed-Model-ANOVA mit den unabhängigen Faktoren Wahrheitsstatus, Vorbereitungszeit und der Aussageform als Messwiederholungsfaktor zeigte einen signifikanten Effekt der Form der Aussage auf deren Eignung für die Inhaltsanalyse. Die Interviews wurden hinsichtlich ihrer Eignung besser beurteilt (M = 5.65, SD = 1.79) als die freien Berichte (M = 4.95, SD = 2.10), F(1,172) = 39.09, p < .001, r = .18. Es ergaben sich keine Wechselwirkungen zwischen der Aussageform und den anderen beiden unabhängigen Faktoren, alle Fs(1,172) ≤ 1.91, ps ≥ .169, fs ≤ .11. Allerdings erschienen wahre (M = 5.75) im Vergleich zu erfundenen Aussagen (M = 4.85) als besser geeignet, F(1,172) = 11.92, p = .001, r = .25. Ebenso variierte die Eignung mit dem Ausmaß an Vorbereitungszeit. Vorbereitete Aussagen ließen sich besser analysieren (M = 5.63), als unvorbereitete (M = 4.97), F(1,172) = 6.41, p = .012, r = .19. Die Wechselwirkung zwischen dem Wahrheitsstatus und der Vorbereitungszeit war jedoch nicht signifikant, F(1,172) = 1.28, p = . 260, f = .10.

Getrennt für die freien Berichte und die Interviews wurde überprüft, ob sich die experimentellen Variationen auf die Aussagelänge auswirkten. Dazu wurden 2 x 2 ANOVAs mit den Faktoren Wahrheitsstatus und Vorbereitungszeit und mit der Anzahl der Wörter als abhängiger Variable durchgeführt. Für die freien Berichte ergab sich ein geringer, jedoch signifikanter Haupteffekt des Wahrheitsstatus, F(1,172) = 8.03, p = .005, r = .20. Wahre Aussagen (M = 857, 95% KI = 769-946) umfassten mehr Wörter als erfundene (M = 678, 95% KI = 589-766). Zudem zeigte sich ein signifikanter Haupteffekt der Vorbereitungszeit, F(1,172) = 19.64, p < .001, r = .31. Vorbereitete Aussagen (M = 908, 95% KI = 819-996) waren umfangreicher als unvorbereitete (M = 627, 95% KI = 539-715). Eine Wechselwirkung zwischen den beiden unabhängigen Variablen auf die Aussagelänge war nicht nachweisbar, F(1,172) = 0.55 , p = .461, f = .06.

Für die Interviews ergab sich ebenfalls ein signifikanter Effekt des Wahrheitsstatus auf die Aussagelänge, $F(1,172) = 5.32$, $p = .022$, $r = .17$. Dieser ließ sich erneut darauf zurückführen, dass wahre Aussagen ($M = 1103$, 95% KI = 992-1214) umfangreicher waren als erfundene ($M = 919$, 95% KI = 808-1030). Die Aussagelänge variierte hingegen nicht signifikant mit dem Ausmaß an Vorbereitungszeit, $F(1,172) = 1.71$, $p = .193$, $r = .10$ (ohne Vorbereitung: $M = 959$, 95% KI = 848-1070, mit Vorbereitung: $M = 1063$, 95% KI = 952-1174). Auch für die Interviews lag keine Wechselwirkung zwischen den variierten Faktoren vor, $F(1,172) = 0.03$, $p = .856$, $f = .00$. Schließlich ergab sich eine substanzielle Korrelation zwischen der Aussagelänge der freien Berichte und der Interviews ($r(174) = .796$, $p < .001$).

Analysen anhand von Pearson-Korrelationen verwiesen zudem auf einen positiven Zusammenhang zwischen der Länge der Aussagen und ihrer Eignung für die inhaltliche Aussageanalyse. Dies galt sowohl für die freien Berichte ($r(174) = .60$, $p < .001$), als auch für die Interviews ($r(174) = .73$, $p < .001$). Es zeigten sich ebenfalls positive Zusammenhänge zwischen der Aussagelänge und dem subjektiven Glaubhaftigkeitsurteil der Raterinnen für die freien Berichte ($r(174) = .64$, $p < .001$) und die Interviews ($r(174) = .74$, $p < .001$). Mit zunehmendem Umfang waren höhere Glaubhaftigkeitsurteile festzustellen. Die Valenz der Ereignisse korrelierte hingegen nicht mit der Aussagelänge, weder für die freien Berichte ($r(174) = -.05$, $p = .525$) noch für die Interviews ($r(174) = -.05$, $p = .488$).

Zusammenfassend zeigten die Voranalysen, dass wahre Aussagen umfangreicher waren als erfundene. Für die freien Berichte beeinflusste zudem die Variation der Vorbereitungszeit die Anzahl der Wörter. Die ARJS-Aussageanalyse erfolgte anhand abgestufter Beurteilungsskalen. Es wird davon ausgegangen, dass diese Form der Beurteilung gegenüber Variationen in der Aussagelänge relativ robust ist (vgl. Granhag et al., 2006; Strömwall et al., 2004). Daher wurde die Aussagelänge bei den nachfolgenden Analysen nicht kontrolliert.

Im Anhang 3a finden sich jedoch die Ergebnisse entsprechender Zusatzanalysen mit der Aussagelänge als Kovariate.

Deskriptive Analysen und interne Konsistenz der ARJS

Die Itembeurteilungen wurden gemäß der Skalenstruktur der ARJS zusammengefasst (vgl. Tabelle 3.5). Die internen Konsistenzen waren für die meisten Skalen zufrieden stellend. Unzureichende interne Konsistenzen, beispielsweise für die Skalen Realismus und logische Struktur sowie Sinneseindrücke, ließen sich jedoch auf eine mangelnde Variabilität der Merkmale im Stimulusmaterial zurückführen.

Tabelle 3.6 stellt die Mittelwerte, Standardabweichungen, Minima und Maxima für die 13 ARJS-Skalen dar. Diese basieren auf den über alle Aussagen und Raterinnen gemittelten Beurteilungen. Dabei lagen für die freien Berichte jeweils zwei und für die Interviews vier unabhängige Beurteilungen vor. Sowohl für die freien Berichte als auch für die Interviews zeigten sich deutliche Deckeneffekte hinsichtlich der Skala Realismus und logische Struktur. Hingegen lagen in dem Stimulusmaterial kaum Informationen zu Sinneseindrücken vor. Für beide Aussageformen wiesen auch die Skalen Klarheit und Lebendigkeit sowie Komplikationen und ungewöhnliche Details deutlich erhöhte bzw. reduzierte Mittelwerte auf. Jedoch lag für diese Skalen zumindest eine gewisse Variabilität zwischen den Aussagen vor.

Inter-Rater-Reliabilität der ARJS

Die Inter-Rater-Reliabilitäten wurden getrennt für die freien Berichte und die Interviews anhand von Produkt-Moment-Korrelationen ermittelt. Dadurch bleibt das Intervallskalenniveau der verwendeten siebenstufigen Ratingskala erhalten. Die Produkt-Moment-Korrelation stellt ein justiertes Reliabilitätsmaß für zwei Rater dar. Die Beurteilungen eines Raters werden an seinem individuellen Mittelwert z-standardisiert, so dass sich Mittelwertsunterschiede zwischen den Ratern nicht korrelationsmindernd auswirken. Auch Varianzunterschiede zwischen den Ratern

Tabelle 3.5

Skalen und Items der Aberdeen Report Judgment Scales (ARJS) und deren interne Konsistenz im freien Bericht und im Interview

Skalen	Items	Cronbach's Alpha: Freie Berichte	Cronbach's Alpha: Interviews
(01) Realismus und logische Struktur[a]	(01) Unplausible Details (–) (02) Realismus (03) Widersprüche (04) Rekonstruierbare Struktur*	.33	.44
(02) Klarheit und Lebendigkeit[a]	(05) Klarheit (06) Visuelle Details (07) Lebendigkeit (08) Spontane Organisation	.71	.79
(03) Details[a]	(09) Details im Hauptereignis (10) Details im Nebenereignis (11) Präzise Details (12) Überflüssige Details[a]	.87	.91
(04) Räumliche Details[a]	(13) Räumliche Details gesamt (14) Details zu Ort und Umgebung (15) Anordnung von Gegenständen (16) Anordnung von Personen	.75	.82
(05) Zeitliche Details[a]	(17) Zeitliche Details gesamt (18) Angaben zur Jahreszeit (19) Angaben zum Jahr (20) Angaben zu Tag oder Datum (21) Angaben zur Tages- oder Uhrzeit	.73	.63

Tabelle 3.5 (Fortsetzung)

Skalen	Items	Cronbach's Alpha: Freie Berichte	Cronbach's Alpha: Interviews
(06) Sinneseindrücke[a]	(22) Geräusche (23) Gerüche (24) Tastempfindungen (25) Geschmacksempfindungen*	.13	.31
(07) Emotionen und Gefühle[a]	(26) Emotionen und Gefühle (27) Intensive Gefühle (28) Emotionstypen (29) Gefühlsverlauf	.82	.90
(08) Gedanken[a]	(30) Gedanken (31) Präzise Gedanken (32) Schlussfolgernde Prozesse	.78	.83
(09) Memorieren und Gedächtnis[a]	(33) Sicherheit der eigenen Angaben (–) (34) Wiederholtes Nachdenken (35) Wiederholtes Erzählen (36) Ereignisse vorher (37) Ereignisse nachher (38) Unterstützende Erinnerungen	.61	.63
(10) Nonverbale und verbale Interaktionen[a]	(39) Nonverbale Interaktionen (40) Präzise nonverbale Interaktionen (41) Verbale Interaktionen (42) Präzise verbale Interaktionen (43) Gedanken und Gefühle anderer Personen	.72	.74

Tabelle 3.5 (Fortsetzung)

Skalen	Items	Cronbach's Alpha: Freie Berichte	Cronbach's Alpha: Interviews
(11) Komplikationen und ungewöhnliche Details	(44) Komplikationen[a] (45) Ungewöhnliche Details[a]	.45	.52
(12) Fehler und sozial Unerwünschtes	(46) Korrekturen oder Präzisierungen[a] (47) Erinnerungslücken[a] (48) Mangel an sozialer Erwünschtheit[a]	.49	.60
(13) Persönliche Signifikanz[a]	(49) Persönliche Bedeutung (50) Scheinbar ernsthafte Folgen (51) Tatsächlich ernsthafte Folgen (52) Persönlichkeit des Erzählers	.53	.66

Anm. N = 176. (–) Items werden umgepolt. *Wegen fehlender Varianz gingen Items 4 und 25 bei den freien Berichten nicht in die Analyse der internen Konsistenz mit ein. [a] = vergleichbare Glaubhaftigkeitsmerkmale sind in der ARJS-STV-G integriert.

Tabelle 3.6

Mittelwerte und Standardabweichungen der 13 ARJS-Skalen für die freien Berichte (N =176) und Interviews (N =176)

	Freie Berichte				Interviews			
Skalen	M	SD	Min	Max	M	SD	Min	Max
Realismus und logische Struktur	6.98	0.11	6.25	7.00	6.96	0.10	6.44	7.00
Klarheit und Lebendigkeit	5.27	0.97	2.63	7.00	5.61	0.86	2.69	6.88
Details	3.93	1.31	1.50	6.88	4.25	1.16	1.69	6.88
Räumliche Details	2.60	1.06	1.00	5.38	3.39	1.17	1.00	6.44
Zeitliche Details	3.36	1.54	1.00	7.00	4.44	1.40	1.35	7.00
Sinneseindrücke	1.25	0.48	1.00	3.25	1.31	0.50	1.00	3.44
Emotionen und Gefühle	3.86	1.33	1.00	6.75	3.73	1.26	1.00	6.56
Gedanken	4.07	1.77	1.00	7.00	3.96	1.74	1.00	7.00
Memorieren und Gedächtnis	2.48	0.98	1.00	5.60	2.53	0.89	1.10	4.95
Nonverbale und verbale Interaktionen	2.90	1.20	1.00	6.00	2.96	1.15	1.00	6.30
Komplikationen und ungewöhnliche Details	1.85	1.02	1.00	4.75	1.95	0.90	1.00	4.38
Fehler und sozial Unerwünschtes	1.96	0.96	1.00	6.25	2.53	0.99	1.00	5.31
Persönliche Signifikanz	3.62	1.11	1.25	6.00	3.54	1.07	1.50	5.50

Anm. Alle Skalenwerte resultierten aus den 7-stufigen Beurteilungen der entsprechenden Einzelitems.

werden durch die Produkt-Moment-Korrelation nicht abgebildet. Da für nachfolgende Analysen die über alle Raterinnen gemittelten Urteile verwendet wurden, wurde zudem anhand der Spearman-Brown-Formel $r_{SP(n)} = (n*r)/(1+(n-1)*r)$ die Reliabilität der Gruppenurteile geschätzt (Rosenthal, 1995). Dabei bezieht sich r auf die mittlere Korrelation zwischen den Einzelratings und n auf die Anzahl der Raterinnen, die zum Gruppenurteil beitragen.

Freie Berichte

Je zwei Raterinnen beurteilten jeweils die Hälfte der freien Berichte, so dass für jeden freien Bericht zwei unabhängige Beurteilungen vorlagen. Die Korrelationen zwischen den Urteilen der beiden Raterinnen, die dasselbe Stimulusmaterial beurteilten, sind in Tabelle 3.7 aufgeführt. Um die Inter-Rater-Reliabilität für das gesamte Stimulusmaterial zu ermitteln, wurden die Korrelationen für beide Raterinnenpaare zunächst in Fisher's Z-Werte umgerechnet und dann gemittelt. Die resultierenden Werte wurden dann in r zurück transformiert.

Die gemittelten Korrelationen verwiesen auf eine äußerst reliable Erfassung der Skalen zeitliche Details, Details sowie nonverbale und verbale Interaktionen. Auch für die übrigen Skalen wurden meist zufrieden stellende Inter-Rater-Reliabilitäten erzielt. Reliabilitätsmaße setzen jedoch eine deutliche Merkmalsvarianz im Stimulusmaterial voraus. Daher ist es nicht verwunderlich, dass für die Skalen, bei denen Boden- und Deckeneffekte vorlagen, nur geringe Korrelationen nachgewiesen wurden. Wenn beide Raterinnen konsistent hohe oder geringe Beurteilungen abgeben, liegt kaum Varianz vor, die sich korrelationsstatistisch aufklären ließe. Ein besonders deutlicher Deckeneffekt ergab sich hinsichtlich der Skala Realismus und logische Struktur (vgl. Tabelle 3.6). Da eine Raterin für alle freien Berichte ein Rating von 7 vergab, ließ sich für diese Skala kein korrelativer Zusammenhang berechnen. Doch auch die Beurteilungen der übrigen Raterinnen variierten kaum, so dass der für diese Skala resultierende Korrelationskoeffizient nicht aussagekräftig ist.

Tabelle 3.7
Mittelwerte, Standardabweichungen, Inter-Rater-Reliabilitäten (Pearson-Korrelationen) der ARJS-Urteile für die freien Berichte (B01.01 - B11.16)

	Raterin W		Raterin Y			Raterin X		Raterin Z				
Skalen	M_W	SD_W	M_Y	SD_Y	r_{WY}	M_X	SD_X	M_Z	SD_Z	r_{XZ}	M von r	$r_{SP(2)}$
Realismus und logische Struktur	6.97	0.18	6.98	0.11	.02	6.95	0.25	7.00	0.00	-	-	-
Klarheit und Lebendigkeit	4.19	1.29	5.46	0.71	.49	5.60	1.01	5.81	0.91	.57	0.53	0.69
Details	3.52	1.36	3.62	1.55	.80	4.32	1.26	4.27	1.22	.71	0.75	0.86
Räumliche Details	1.84	0.98	2.98	1.20	.65	2.53	1.23	3.03	1.15	.69	0.67	0.80
Zeitliche Details	2.95	1.29	3.53	1.61	.86	3.31	1.62	3.63	1.79	.93	0.90	0.95
Sinneseindrücke	1.11	0.35	1.31	0.52	.59	1.32	0.64	1.26	0.53	.75	0.68	0.81
Emotionen und Gefühle	3.91	1.52	4.59	1.43	.53	3.30	1.58	3.63	1.23	.60	0.57	0.73
Gedanken	3.29	1.83	4.59	2.03	.59	4.30	1.99	4.12	1.90	.78	0.69	0.82
Memorieren und Gedächtnis	2.48	0.93	2.13	0.98	.68	2.67	1.26	2.62	1.05	.64	0.66	0.79
Nonverbale und verbale Interaktionen	2.69	1.18	3.15	1.46	.85	3.04	1.28	2.74	1.14	.78	0.82	0.90
Komplikationen und ungewöhnliche Details	1.19	0.55	2.18	1.39	.45	2.31	1.62	1.72	0.94	.58	0.52	0.69
Fehler und sozial Unerwünschtes	1.36	0.69	2.16	1.23	.43	2.21	1.15	2.13	1.14	.67	0.57	0.72
Persönliche Signifikanz	4.00	1.33	3.39	1.12	.69	3.65	1.40	3.45	1.02	.65	0.67	0.80

Anm. N = 176. 7-stufige Ratingskalen. Jedes Raterinnenpaar beurteilte n = 88 Aussagen, Raterinnenpaar W-Y beurteilte die ein Hälfte (B01.01 - B06.08), Rateinnenpaar X-Z die andere Hälfte der Berichte (B06.09 - B11.16). M von r basiert auf den unabhängigen Beurteilungen durch zwei Raterinnen. $r_{SP(2)}$: Spearman-Brown-Korrektur für zwei Rater.

Allerdings zeigten Zusatzanalysen, bei denen nur identische Urteile als übereinstimmend aufgefasst wurden, dass eine hohe prozentuale Übereinstimmung von 95% vorlag. Demnach ist trotz korrelationsstatistisch nicht nachweisbarer Reliabilität der Urteile eine hohe Übereinstimmung der Beurteilungen festzustellen.

Für nachfolgende Analysen der freien Berichte wurden die über jeweils zwei Raterinnen gemittelten ARJS-Beurteilungen verwendet. Daher erschien es sinnvoll neben der Reliabilität der Einzelratings abzuschätzen, wie reliabel das gemeinsame Urteil zweier Rater ausfällt. Analysen anhand der Spearman-Brown-Formel zeigten, dass die gemeinsamen Urteile für alle Skalen deutlich reliabler sind als die Einzelratings. Ausgenommen der Skalen Realismus und logische Struktur sowie Komplikationen und ungewöhnliche Details waren alle Gruppenurteile auf Skalenebene als zufrieden stellend anzusehen ($.69 \leq r_{SP(2)} \leq .95$).

Interviews

Für jedes Interview lagen vier unabhängige Beurteilungen vor. Um die Inter-Rater-Reliabilität der Einzelratings zu bestimmen, wurden zunächst die Beurteilungen aller Raterinnen paarweise miteinander korreliert. Die resultierenden Korrelationen wurden anschließend in Fisher's Z-Werte transformiert, dann gemittelt und in r zurück transformiert (Tabelle 3.8). Auch für die Interviews zeigten sich die höchsten Korrelationen hinsichtlich der Skalen zeitliche Details, Details sowie nonverbale und verbale Interaktionen. Doch auch die Skala Gedanken wurde bei den Interviews äußerst reliabel erfasst. Erneut zeigten sich geringere Inter-Rater-Reliabilitäten für die Skalen, bei denen Boden- und Deckeneffekte nachgewiesen wurden. Die anhand der Spearman-Brown-Formel geschätzten Reliabilitäten des Gruppenurteils von vier Raterinnen waren für alle Skalen als hoch zu werten ($.65 \leq r_{SP(4)} \leq .96$). Lediglich die Skala Realismus und logische Struktur ist hiervon auszuschließen. Aufgrund der fehlenden Variabilität dieses Merkmals im Stimulusmaterial war der Korrelationskoeffizient nicht dazu

Tabelle 3.8

Inter-Rater-Reliabilitäten (Pearson-Korrelationen) der ARJS-Skalen für die Interviews (C01.01 – C11.16)

Skalen	r_{WX}	r_{WY}	r_{WZ}	r_{XY}	r_{XZ}	r_{YZ}	M von r	$r_{SP(4)}$
Realismus und logische Struktur	-.06	-.02	-.02	-.03	-.02	-.01	-.02	.09
Klarheit und Lebendigkeit	.52	.43	.50	.58	.65	.60	.55	.83
Details	.66	.76	.70	.74	.69	.69	.71	.91
Räumliche Details	.62	.57	.68	.72	.75	.72	.68	.89
Zeitliche Details	.83	.86	.79	.87	.88	.84	.84	.96
Sinneseindrücke	.63	.56	.61	.66	.65	.62	.62	.87
Emotionen und Gefühle	.56	.66	.64	.62	.66	.70	.64	.88
Gedanken	.72	.69	.73	.75	.82	.73	.74	.92
Memorieren und Gedächtnis	.65	.57	.61	.50	.63	.64	.60	.86
Nonverbale und verbale Interaktionen	.72	.75	.75	.78	.80	.78	.76	.93
Komplikationen und ungewöhnliche Details	.20	.20	.21	.38	.48	.46	.32	.65
Fehler und sozial Unerwünschtes	.55	.51	.51	.56	.64	.53	.55	.83
Persönliche Signifikanz	.55	.63	.67	.64	.67	.72	.65	.88

Anm. N = 176. Alle vier Raterinnen beurteilten sämtliche Interviews. M von r basiert auf den unabhängigen Beurteilungen durch vier Raterinnen. $r_{SP(4)}$: Spearman-Brown-Korrektur für vier Rater.

geeignet, die Inter-Rater-Reliabilität abzubilden. Es ließ sich jedoch eine hohe prozentuale Übereinstimmung von 94% nachweisen.

Validität der ARJS

Um die Validität der ARJS zu überprüfen, wurden die freien Berichte und die Interviews zunächst getrennt analysiert. Anhand von 2 x 2 MANOVAs wurden der Einfluss des Wahrheitsstatus und der Vorbereitungszeit auf die 13 ARJS-Skalen überprüft. Für die freien Berichte basieren die ARJS-Urteile auf den über zwei, für die Interviews auf den über vier Raterinnen gemittelten Beurteilungen. Zudem wurden Diskriminanzanalysen durchgeführt, um die Klassifikationsgüte der ARJS abzuschätzen. Dazu wurden die 13 ARJS-Skalen als Prädiktorvariablen und der objektive Wahrheitsstatus als Kriterium verwendet. Schließlich wurden die ARJS-Urteile der freien Berichte und der Interviews direkt miteinander verglichen, indem die Aussageform als Messwiederholungsfaktor bei den Analysen berücksichtigt wurde.

Freie Berichte

Es zeigte sich ein multivariat signifikanter Effekt des Wahrheitsstatus auf die Beurteilung der ARJS, Wilks' Lambda = .82, $F(13,160)= 2.80$, $p = .001$, partielles $eta^2 = .19$. Auf univariater Ebene ergaben sich signifikante Mittelwertsunterschiede für vier Skalen. Die ARJS-Merkmale waren erwartungsgemäß häufiger in wahren als in erfundenen Aussagen vorzufinden. Es zeigten sich Effekte mittlerer Größenordnung für die Skalen Details, Klarheit und Lebendigkeit, Emotionen und Gefühle (Tabelle 3.9). Zudem resultierte für die Skala Komplikationen und ungewöhnliche Details ein geringer Effekt. Der Mittelwertsunterschied hinsichtlich der Skala Memorieren und Gedächtnis verfehlte nur knapp statistische Signifikanz. Ausgenommen der Skala persönliche Signifikanz wiesen auch die übrigen Mittelwertsunterschiede in die erwartete Richtung.

Tabelle 3.9

Mittelwertsunterschiede zwischen wahren und erfundenen Aussagen hinsichtlich der 13 ARJS-Skalen

Skalen	Freie Berichte					Interviews				
	$M_{erfunden}$	M_{wahr}	$F(1,172)$	p	r	$M_{erfunden}$	M_{wahr}	$F(1,172)$	p	r
Realismus und logische Struktur	6.97	6.99	1.53	.218	.09	6.96	6.97	0.16	.686	.03
Klarheit und Lebendigkeit	5.01	5.52	12.70	**.000**	**.26**	5.35	5.86	16.73	**.000**	**.30**
Details	3.58	4.28	14.48	**.000**	**.27**	3.87	4.64	21.90	**.000**	**.33**
Räumliche Details	2.55	2.64	0.31	.580	.04	3.27	3.51	1.78	.185	.10
Zeitliche Details	3.17	3.54	2.56	.112	.12	4.24	4.63	3.50	.063	.14
Sinneseindrücke	1.27	1.23	0.36	.552	-.05	1.31	1.31	0.01	.910	.01
Emotionen und Gefühle	3.57	4.14	8.45	**.004**	**.22**	3.44	4.02	9.93	**.002**	**.23**
Gedanken	3.85	4.30	2.84	.094	.13	3.71	4.20	3.59	.060	.14
Memorieren und Gedächtnis	2.34	2.61	3.41	.067	.14	2.36	2.70	6.58	**.011**	**.19**
Nonverbale und verbale Interaktionen	2.78	3.03	1.91	.169	.10	2.74	3.18	6.63	**.011**	**.19**
Komplikationen und ungewöhnliche Details	1.69	2.01	4.31	**.039**	**.15**	1.74	2.15	9.53	**.002**	**.23**
Fehler und sozial Unerwünschtes	1.92	2.01	0.42	.519	.05	2.31	2.75	8.99	**.003**	**.22**
Persönliche Signifikanz	3.73	3.51	1.60	.208	-.10	3.66	3.43	2.03	.156	-.11

Anm. N = 176, 7-stufige Ratingskalen. Positive Effektstärken r signieren eine höhere Ausprägung in wahren Berichten.

Der Haupteffekt der Vorbereitungszeit auf die Beurteilung der ARJS-Skalen war ebenfalls signifikant, Wilks' Lambda = .84, F(13,160)= 2.35, p = .007, partielles eta^2 = .16. Die Ergebnisse der univariaten Analyse sind in Tabelle 3.10 dargestellt. In den Bedingungen mit Vorbereitung ergaben sich signifikant höhere Beurteilungen als ohne Vorbereitung für die Skalen Details, nonverbale und verbale Interaktionen, Memorieren und Gedächtnis sowie zeitliche Details. Dabei lagen Effektstärken in geringer bis mittlerer Größenordnung vor. Zudem zeigte sich ein marginal signifikanter Effekt für die Skala Gedanken. Ausgenommen der Skala Realismus und logische Struktur verwiesen die übrigen Mittelwertsunterschiede ebenfalls auf eine höhere Qualität vorbereiteter im Vergleich zu unvorbereiteten Aussagen. Es zeigte sich keine signifikante Wechselwirkung zwischen dem Wahrheitsstatus und der Vorbereitungszeit, Wilks' Lambda = .91, F(13,160)= 1.19, p = .293, partielles eta^2 = .09.[3]

---

[3] Zusätzlich wurde eine 2 x 2 MANCOVA gerechnet, bei der die Anzahl der Wörter als Kovariate mit einging (vgl. Anhang 3a). Diese verwies ebenfalls auf einen signifikanten Effekt des Wahrheitsstatus auf die Beurteilung der 13 ARJS-Skalen, Wilks' Lambda = 0.85, F(13,159) = 2.11, p = .016, partielles eta^2 = .15. Erneut zeigten sich signifikante Mittelwertsunterschiede für die Skalen Klarheit und Lebendigkeit sowie Details. Zusätzlich ergab sich unter Kontrolle der Anzahl der Wörter ein signifikanter Effekt für die Skala persönliche Signifikanz. Sie war bei erfundenen Aussagen höher ausgeprägt als bei wahren (r = -.18, vgl. Anhang 3a). Ein Haupteffekt der Vorbereitungszeit war unter Einschluss der Kovariaten nicht mehr nachweisbar, Wilks' Lambda = 0.92, F(13,159) = 1.13, p = .338, partielles eta^2 = .09. Auch die Wechselwirkung zwischen dem Wahrheitsstatus und der Vorbereitungszeit war nicht signifikant, Wilks' Lambda = 0.91, F(13,159) = 1.23, p = .263, partielles eta^2 = .09.

Tabelle 3.10

Mittelwertsunterschiede zwischen unvorbereiteten und vorbereiteten freien Berichten hinsichtlich der 13 ARJS-Skalen (B01.01-B11.16)

Skalen	$M_{ohne\ Vorb.}$	$M_{mit\ Vorb.}$	$F(1,172)$	p	r
Realismus und logische Struktur	6.99	6.97	1.53	.218	-.09
Klarheit und Lebendigkeit	5.16	5.38	2.40	.123	.11
Details	3.59	4.27	3.56	**.000**	**.26**
Räumliche Details	2.46	2.73	2.94	.088	.13
Zeitliche Details	3.07	3.64	6.13	**.014**	**.18**
Sinneseindrücke	1.24	1.25	0.01	.905	.01
Emotionen und Gefühle	3.69	4.03	3.01	.084	.13
Gedanken	3.81	4.33	3.83	.052	.15
Memorieren und Gedächtnis	2.29	2.66	6.60	**.011**	**.19**
Nonverbale und verbale Interaktionen	2.65	3.16	8.36	**.004**	**.21**
Komplikationen und ungewöhnliche Details	1.72	1.98	3.03	.084	.13
Fehler und sozial Unerwünschtes	1.85	2.07	2.34	.128	.12
Persönliche Signifikanz	3.59	3.65	0.10	.749	.02

Anm. N = 176, 7-stufige Ratingskalen. Positive Effektstärken r signieren eine höhere Ausprägung in vorbereiteten Berichten.

Eine multiple Diskriminanzanalyse erlaubte es, den objektiven Wahrheitsstatus anhand der 13 ARJS-Skalen signifikant vorherzusagen, Wilks' Lambda = .82, chi^2 (13) = 32.98, p = .002. Insgesamt ließen sich 71.0% der freien Berichte richtig klassifizieren. Die Klassifikationsgüte war für erlebnisbasierte Aussagen höher (75.0%) als für erfundene (67.0%).

Interviews

Für die Interviews ergab sich ebenfalls ein multivariat signifikanter Effekt des Wahrheitsstatus auf die ARJS-Beurteilungen, Wilks' Lambda = .78, F(13,160)= 3.53, p < .001, partielles eta^2 = .22. Dieser war univariat auf signifikante Mittelwertsunterschiede für sieben Skalen zurückzuführen (Tabelle 3.9). Dabei wiesen wahre Aussagen eine höhere Aussagequalität auf als erfundene. In Übereinstimmung mit den Befunden für die freien Berichte zeigten sich signifikante Effekte des Wahrheitsstatus auf die Skalen Klarheit und Lebendigkeit, Details, Emotionen und Gefühle, Komplikationen und ungewöhnliche Details. Zudem enthielten wahre Aussagen im Rahmen der Interviews mehr Verweise auf Fehler und sozial Unerwünschtes, Memorieren und Gedächtnisprozesse sowie nonverbale und verbale Interaktionen als erfundene. Die Effektstärken lagen überwiegend im mittleren bis hohen Bereich. Zudem wurden wahre Aussagen marginal höher hinsichtlich der Skala Gedanken beurteilt. Auch die übrigen Skalen, ausgenommen der persönlichen Signifikanz, bildeten erwartungsgemäß eine höhere Qualität bei wahren im Vergleich zu erfundenen Aussagen ab.

Im Gegensatz zu den freien Berichten ergab sich für die Interviews kein multivariat signifikanter Haupteffekt der Vorbereitungszeit auf die ARJS-Beurteilungen, Wilks' Lambda = .96, F(13,160)= .53, p = .903, partielles eta^2 = .04. Die Interaktion zwischen dem Wahrheitsstatus und der Vorbereitungszeit war

ebenfalls nicht signifikant, Wilks' Lambda = 0.97, F(13,160) = 0.44, p = .951, partielles eta^2 = .04.[4]

Auch für die Interviews erlaubten die ARJS-Skalen den tatsächlichen Wahrheitsstatus überzufällig gut vorherzusagen, Wilks' Lambda = .78, chi^2 (13) = 41.62, p < .001. Die multiple Diskriminanzfunktion erzielte insgesamt 72.2% richtige Klassifikationen. Wahre und erfundene Aussagen ließen sich dabei ähnlich gut klassifizieren, 72.7% und 71.6% respektive.

Vergleich der ARJS-Beurteilungen beider Aussageformen

Um zu überprüfen, ob die Beurteilungen der 13 ARJS-Skalen in Abhängigkeit von der Aussageform systematisch variierten, wurde eine weitere 2 x 2 (x 2) MANOVA gerechnet. Als unabhängige Between-Subjects-Faktoren gingen der Wahrheitsstatus und die Vorbereitungszeit mit ein. Die Variation der Aussageform (freie Berichte versus Interviews) wurde als Messwiederholungsfaktor aufgefasst.

Auch bei gemeinsamer Betrachtung beider Aussageformen zeigte sich ein signifikanter Haupteffekt des Wahrheitsstatus auf die ARJS-Urteile, Wilks' Lambda

---

[4] Der signifikante Haupteffekt des Wahrheitsstatus ließ sich durch eine 2 x 2 MANCOVA mit der Anzahl der Wörter als Kovariate replizieren, Wilks' Lambda = 0.78, F(13,159) = 3.48, p < .001, partielles eta^2 = .22. Erneut ergaben sich erwartungsgemäße Mittelwertsunterschiede für die Skalen Details, Klarheit und Lebendigkeit, Komplikationen und ungewöhnliche Details, Fehler und sozial Unerwünschtes sowie Emotionen und Gefühle. Zudem erzielten erfundene Aussagen im Vergleich zu wahren signifikant höhere Bewertungen hinsichtlich der persönlichen Signifikanz des Ereignisses (r = -.18, vgl. Anhang 3a). Es ergab sich auch bei Kontrolle der Aussagelänge weder ein Haupteffekt der Vorbereitungszeit, Wilks' Lambda = 0.97, F(13,159) = 0.42, p = .960, partielles eta^2 = .03, noch eine Wechselwirkung zwischen den beiden unabhängigen Faktoren, Wilks' Lambda = 0.97, F(13,159) = 0.44, p = .953, partielles eta^2 = .04.

= 0.78, F(13,160) = 3.53, p < .001, partielles eta^2 = .22. Wahre Aussagen erhielten signifikant höhere Beurteilungen für acht Skalen als erfundene. Starke bis mittlere Effekte ergaben sich für die Skalen Details (r = .32), Klarheit und Lebendigkeit (r = .31), Emotionen und Gefühle (r = .25), sowie Komplikationen und ungewöhnliche Details (r = .21). Des Weiteren wurden die Skalen Memorieren und Gedächtnis (r = .18), nonverbale und verbale Interaktionen (r = .16), Fehler und sozial Unerwünschtes (r = .16) sowie zeitliche Details (r = .15) in Abhängigkeit vom Wahrheitsstatus unterschiedlich beurteilt. Ein Haupteffekt der Vorbereitungszeit war über beide Aussageformen hinweg nicht nachweisbar, Wilks' Lambda = 0.92, F(13,160) = 1.09, p = .376, partielles eta^2 = .08. Auch die Wechselwirkung zwischen der Vorbereitungszeit und dem Wahrheitsstatus war nicht signifikant, Wilks' Lambda = 0.94, F(13,160) = 0.74, p = .717, partielles eta^2 = .06.

Zudem ergab sich ein signifikanter Effekt der Aussageform, Wilks' Lambda = 0.42, F(13,160) = 16.96, p < .001, partielles eta^2 = .58. Tabelle 3.11 ist zu entnehmen, dass sechs Skalen bei den Interviews höhere Bewertungen erzielten als bei den freien Berichten. Die stärksten Effekte zeigten sich für die Skalen räumliche und zeitliche Details, Fehler und sozial Unerwünschtes, gefolgt von den Skalen Klarheit und Lebendigkeit sowie Details. Die Beurteilung der Skala Sinneseindrücke variierte in Abhängigkeit von der Aussageform in entsprechender Richtung. Hingegen erzielte die Skala persönliche Signifikanz geringfügig höhere Beurteilungen bei den freien Berichten als bei den Interviews. Dieser Unterschied verfehlte jedoch knapp statistische Signifikanz.

Eine multivariate Wechselwirkung zwischen der Aussageform und dem Wahrheitsstatus war nicht festzustellen, Wilks' Lambda = 0.94, F(13,160) = 0.74, p = .717, partielles eta^2 = .06.

Tabelle 3.11

Mittelwertsunterschiede zwischen freien Berichten (B01.01-B11.16) und Interviews (C01.01-C11.16) hinsichtlich der 13 ARJS-Skalen

Skalen	$M_{Berichte}$	$M_{Interviews}$	$F(1,172)$	p	r
Realismus und logische Struktur	6.98	6.96	1.63	.204	-.06
Klarheit und Lebendigkeit	5.27	5.61	29.88	**.000**	**.18**
Details	3.93	4.25	28.45	**.000**	**.13**
Räumliche Details	2.60	3.39	92.38	**.000**	**.34**
Zeitliche Details	3.36	4.44	95.76	**.000**	**.34**
Sinneseindrücke	1.25	1.31	3.92	**.049**	**.06**
Emotionen und Gefühle	3.86	3.73	1.81	.181	-.05
Gedanken	4.07	3.96	1.31	.254	-.03
Memorieren und Gedächtnis	2.48	2.53	1.08	.300	.03
Nonverbale und verbale Interaktionen	2.91	2.96	0.87	.352	.02
Komplikationen und ungewöhnliche Details	1.85	1.95	2.43	.121	.05
Fehler und sozial Unerwünschtes	1.96	2.53	56.86	**.000**	**.28**
Persönliche Signifikanz	3.62	3.54	3.86	.051	-.04

Anm.: N = 176 freie Berichte und N = 176 Interviews. Positive Effektstärken r signieren eine höhere Ausprägung bei den Interviews.

Allerdings zeigte sich auf univariater Ebene ein signifikanter Interaktionseffekt hinsichtlich der Skala Fehler und sozial Unerwünschtes, $F(1,172) = 5.40$, $p = .021$, $f = .18$. Für die Interviews ergaben sich Mittelwertsunterschiede zwischen wahren ($M = 2.75$) und erfundenen Aussagen ($M = 2.31$), $F(1,174) = 9.09$, $p = .003$, $r = .22$, die für die freien Berichte nicht feststellbar waren ($M = 2.01$ und $M = 1.92$, respektive), $F(1,174) = 0.42$, $p = .520$, $r = .05$. Die Wechselwirkung ließ sich jedoch wie bereits erwähnt nicht multivariat absichern.

Des Weiteren zeigte sich eine multivariat signifikante Wechselwirkung zwischen der Aussageform und der Vorbereitungszeit, Wilks' Lambda = 0.82, $F(13,160) = 2.70$, $p = .002$, partielles $eta^2 = .18$. Univariat ergaben sich signifikante Unterschiede für die Skalen Details, Gedanken sowie Memorieren und Gedächtnisprozesse, alle $Fs(1,172) \geq 4.54$, $ps \leq .034$, $fs \geq .16$. Während die Vorbereitungszeit bei den freien Berichten noch einen Einfluss zeigte, war dieser nach einer Woche bei den Interviews nicht mehr festzustellen. Wenn die freien Berichte vorbereitet wurden, erhielten sie höhere Beurteilungen hinsichtlich der Skalen Details ($M = 4.27$) sowie Memorieren und Gedächtnis ($M = 2.66$), als wenn sie unvorbereitet erfolgten, $M = 3.59$ und $M = 2.29$ respektive, beide $Fs(1,174) \geq 6.55$, $ps \leq .011$, $r \geq .19$. Auch für die Skala Gedanken waren marginal signifikante Mittelwertsunterschiede zwischen vorbereiteten ($M = 4.33$) und unvorbereiteten ($M = 3.81$) freien Berichten feststellbar, $F(1,174) = 3.80$, $p = .053$, $r = .15$. Für die Interviews hingegen wurden vorbereitete und unvorbereitete Aussagen hinsichtlich der Skalen Details ($M = 4.36$ vs. $M = 4.14$), Gedanken ($M = 4.00$ vs. $M = 3.92$) sowie Memorieren und Gedächtnis ($M = 2.55$ vs. $M = 2.51$) ähnlich beurteilt, alle $Fs(1,174) \leq 1.60$, $ps \geq .207$, $rs \leq .10$. Eine dreifache Wechselwirkung zwischen der Aussageform und den beiden Between-Subjects-Faktoren Wahrheitsstatus und Vorbereitungszeit war nicht festzustellen, Wilks' Lambda = 0.93, $F(13,160) = 0.86$, $p = .592$, partielles $eta^2 = .07$.

Analysen zur Valenz der Ereignisse

Die Raterinnen beurteilten alle Aussagen hinsichtlich ihrer Valenz auf einer 7–stufigen Skala. Die Versuchspersonen beschrieben im Rahmen der freien Berichte und der Interviews jeweils dieselben Ereignisse. Daher erschien es sinnvoll, die Valenzurteile für beide Aussageformen zusammenzufassen. Für jeden freien Bericht lagen zwei und für jedes Interview vier unabhängige Valenzurteile vor. Diese sechs Urteile wiesen substanzielle Interkorrelationen auf ($.62 \leq r \leq .83$) und wurden daher gemittelt (Cronbach's Alpha = .93). Die experimentelle Manipulation des objektiven Wahrheitsstatus hatte keinen signifikanten Einfluss auf die Valenz der Ereignisse, $F(1,174) = 1.63$, $p = .203$, $r = .10$.

Für varianzanalytische Untersuchungen war es zudem erforderlich, einen binären Indikator der Ereignisvalenz abzuleiten. Dies erfolgte über eine Dichotomisierung am Median (Mdn = 2.67), die in einer Gleichverteilung von Ereignissen mit verhältnismäßig positiver ($n = 88$, $M = 4.13$, $SD = 1.42$) und negativer Valenz ($n = 88$, $M = 2.11$, $SD = 0.45$) resultierte. Für positive und negative Ereignisse lagen jeweils gleich viele wahre und erfundene Aussagen vor. Getrennt für die ARJS-Beurteilungen der freien Berichte und der Interviews wurden 2 x 2 MANOVAs mit der Ereignisvalenz und dem Wahrheitsstatus als unabhängigen Faktoren berechnet. Die Befunde sind in Tabelle 3.12 dokumentiert.

Für die freien Berichte ergab sich ein signifikanter Haupteffekt der Valenz des Ereignisses, Wilks' Lambda = 0.75, $F(13,160) = 4.19$, $p < .001$, partielles $eta^2 = .25$. Aussagen über negative Ereignisse zeigten für sechs Skalen eine signifikant höhere Qualität als Aussagen über positive Ereignisse. Dies galt für die Skalen Memorieren und Gedächtnis, persönliche Signifikanz, Klarheit und Lebendigkeit, räumliche Details, Fehler und sozial Unerwünschtes sowie nonverbale und verbale Interaktionen. Eine Wechselwirkung zwischen der Valenz der Ereignisse und dem Wahrheitsstatus war nicht nachweisbar, Wilks' Lambda = 0.90, $F(13,160) = 1.45$, $p = .143$, partielles $eta^2 = .11$.

Tabelle 3.12

Mittelwertsunterschiede zwischen positiven und negativen Ereignissen hinsichtlich der 13 ARJS-Skalen

Skalen	Freie Berichte					Interviews				
	$M_{negativ}$	$M_{positiv}$	F(1,172)	p	r	$M_{negativ}$	$M_{positiv}$	F(1,172)	p	r
Realismus und logische Struktur	6.98	6.97	0.55	.461	.06	6.97	6.95	1.50	.223	-.09
Klarheit und Lebendigkeit	5.43	5.10	5.56	**.020**	**.18**	5.67	5.54	1.04	.309	.08
Details	4.04	3.82	1.37	.243	.09	4.35	4.15	1.47	.226	.09
Räumliche Details	2.79	2.41	5.87	**.016**	**.18**	3.38	3.40	0.02	.877	-.01
Zeitliche Details	3.40	3.31	0.15	.700	.03	4.16	4.71	7.19	**.008**	**-.20**
Sinneseindrücke	1.29	1.21	1.15	.284	.08	1.32	1.29	0.16	.690	.03
Emotionen und Gefühle	3.96	3.75	1.20	.275	.08	3.86	3.61	1.88	.173	.10
Gedanken	4.04	4.10	0.06	.814	-.02	4.12	3.80	1.52	.219	.09
Memorieren und Gedächtnis	2.79	2.16	20.12	**.000**	**.32**	2.82	2.24	21.68	**.000**	**.33**
Nonverbale und verbale Interaktionen	3.11	2.70	5.50	**.020**	**.18**	3.27	2.66	13.77	**.000**	**.27**
Komplikationen und ungewöhnliche Details	1.84	1.85	0.00	.956	-.00	1.91	1.98	0.26	.610	-.04
Fehler und sozial Unerwünschtes	2.13	1.79	5.62	**.019**	**.18**	2.59	2.46	0.73	.393	.07
Persönliche Signifikanz	3.89	3.35	11.06	**.001**	**.25**	3.78	3.30	9.17	**.003**	**.22**

Anm. n = 88 negative und n = 88 positive Ereignisse. Positive Effektstärken r signieren eine höhere Ausprägung in Aussagen über negative Ereignisse.

Auf der Grundlage des vorliegenden Stichprobenumfangs, der empirisch ermittelten Effektstärke eta^2 und einem Alpha-Risiko von alpha = .05 wurde zudem die Teststärke 1-beta berechnet. Die Möglichkeit, dass ein tatsächlich vorliegender Unterschied nicht aufgedeckt wurde, ließ sich mit ausreichender Sicherheit ausschließen, 1-beta = .79.

Für die Interviews ließ sich der multivariat signifikante Effekt der Valenz, Wilks' Lambda = 0.73, $F(13,160) = 4.63$, $p < .001$, partielles eta^2 = .27, auf vier Skalen zurückführen. Erneut gingen negative Ereignisse mit einer höheren Beurteilung der Skalen Memorieren und Gedächtnis, nonverbale und verbale Interaktionen sowie persönliche Signifikanz einher als positive. Im Gegensatz dazu wurden bei positiven Ereignissen signifikant mehr zeitliche Details geschildert als bei negativen.

Die Wechselwirkung zwischen der Valenz der Ereignisse und dem Wahrheitsstatus verfehlte auch für die Interviews statistische Signifikanz, Wilks' Lambda = 0.88, $F(13,160) = 1.74$, $p = .058$, partielles eta^2 = .12. Auf univariater Ebene zeigten sich jedoch signifikante Effekte für die beiden Skalen räumliche Details ($F(1,172) = 5.24$, $p = .023$, $f = .18$) und Gedanken ($F(1,172) = 4.51$, $p = .035$, $f = 16$). Bei positiven Ereignissen erhielten wahre Aussagen erwartungsgemäß höhere Beurteilungen als erfundene. Der einfache Haupteffekt des Wahrheitsstatus war sowohl für räumliche Details ($M = 3.72$ vs. $M = 3.09$, $F(1,172) = 6.61$, $p = .011$, $r = .19$) als auch für Gedanken ($M = 4.32$ vs. $M = 3.28$, $F(1,172) = 8.19$, $p = .005$, $r = .21$) signifikant. Im Gegensatz dazu ergaben sich für negative Ereignisse keine einfachen Haupteffekte. Wahre ($M = 3.29$) und erfundene Aussagen ($M = 3.46$) erhielten vergleichbare Beurteilungen hinsichtlich der Skala räumliche Details, $F(1,172) = 0.44$, $p = .507$, $r = -.05$. Die Skala Gedanken wurde ebenfalls bei wahren ($M = 4.09$) und erfundenen Aussagen ($M = 4.14$) ähnlich beurteilt, $F(1,172) = 0.02$, $p = .889$, $r = -.01$. Wie bereits erwähnt, manifestierten sich diese Wechselwirkungen allerdings nicht auf multivariater Ebene. Mit einer Teststärke von 1-beta = .87 kann erneut davon ausgegangen werden, dass bei

dem vorliegenden Stichprobenumfang ein tatsächlich vorliegender multivariater Unterschied auch statistische Signifikanz erzielt hätte.

Zusammenfassend erzielten Aussagen, deren Valenz negativ eingeschätzt wurde, höhere ARJS-Beurteilungen als Aussagen, deren Valenz als vergleichsweise positiv erachtet wurde. Die ARJS differenzierten jedoch unabhängig von der Valenz gleich gut zwischen wahren und erfundenen Aussagen. Dies schlug sich auch in den subjektiven Urteilen der Raterinnen zur Eignung der Aussagen für die inhaltliche Aussageanalyse nieder. Diese wiesen keinen Zusammenhang zur Valenz der Ereignisse auf, $r(176) = -.03$, $p = .668$, für die freien Berichte, und $r(176) = .00$, $p = .967$, für die Interviews.

Diskussion

Ziel der vorliegenden Untersuchung war es, die Inter-Rater-Reliabilität und Validität der ARJS erneut zu überprüfen. Des Weiteren sollte geklärt werden, ob die Gelegenheit zur Vorbereitung und die Valenz des geschilderten Ereignisses die Validität der ARJS moderieren. Explorativ wurde auch überprüft, ob sich die Qualität wahrer und erfundener Aussagen gleichermaßen unterscheidet, wenn diese frei berichtet bzw. im Rahmen eines Interviews geschildert wurden. Die ARJS-Analyse von jeweils 176 transkribierten freien Berichten und Interviews zu persönlich bedeutsamen Lebensereignissen wurde durch vier trainierte Raterinnen vorgenommen.

Im Folgenden wird zunächst auf die Befunde zur Reliabilität der Beurteilungen eingegangen. Anschließend werden die Ergebnisse zur Validität der ARJS vorgestellt. Dabei werden die Einflüsse des Wahrheitsstatus, der Vorbereitungszeit und der Valenz des geschilderten Ereignisses auf die inhaltliche Aussagequalität dargestellt. Schließlich werden forschungs- und anwendungsbezogene Implikationen der vorliegenden Befunde diskutiert.

Reliabilität der ARJS

Die meisten ARJS-Skalen ließen sich äußerst reliabel erfassen. Dabei ergaben sich die besten Inter-Rater-Reliabilitäten für die Skalen Details, zeitliche Details sowie nonverbale und verbale Interaktionen. Eine gleichermaßen überzeugende Beurteilerübereinstimmung wurde für die Skala Gedanken nachgewiesen. Im Gegensatz dazu erwies sich die Erfassung des RÜ-Merkmals kognitive Operationen oftmals als problematisch (Sporer, 1997a, 1997b; Sporer & Küpper, 1995, Vrij et al., 2004a). Im Rahmen der ARJS wird die Skala Gedanken jedoch als Wahrheitskriterium aufgefasst und entsprechend anders definiert als in der Forschung zum RÜ-Ansatz. Zum einen werden Quantität und Qualität der geschilderten Gedanken getrennt voneinander beurteilt, zum anderen gehen schlussfolgernde Prozesse in die ARJS-Skala mit ein. Auf der Grundlage dieser Definition konnte eine sehr gute Inter-Rater-Reliabilität für die ARJS-Skala Gedanken nachgewiesen werden. Dieser Befund ist mit anderen Untersuchungen zu den ARJS vergleichbar (Barnier et al., 2005; Sporer, 1998; Sporer, Bursch et al., 2000; Sporer & Walther, 2006). Daher lässt sich schlussfolgern, dass die standardisierte und präzise Operationalisierung der ARJS eine zuverlässige Beurteilung dieses Aussagemerkmals gewährleistet.

Vergleicht man die vorliegenden Befunde mit denen der in Tabellen 3.2 und 3.3 dokumentierten CBCA- und RÜ-Forschung, so scheinen die ARJS vergleichbar reliabel zu sein. Allerdings variieren die Befunde verschiedener Untersuchungen zur Inter-Rater-Reliabilität für einzelne RÜ- und CBCA-Merkmalen recht stark. Dies lässt sich zum Teil auf unterschiedliche Beurteilungsmethoden zurückführen. So wurden bei der Verwendung von Häufigkeitsauszählungen oftmals höhere Reliabilitäten erzielt als bei abgestuften Ratingskalen (vgl. auch Vrij, Evans, et al., 2004). Zudem wurden einzelne Merkmale bislang kaum auf ihre Inter-Rater-Reliabilität hin untersucht. Für die CBCA ist dies vor allem darauf zurückzuführen, dass einzelne Merkmale auf das im Labor verwendete Stimulusmaterial nicht anwendbar erscheinen. Für die RÜ-Merkmale wiederum hat sich bislang keine

einheitliche Terminologie durchgesetzt. Infolgedessen ist der Forschungsstand zur Inter-Rater-Reliabilität von Glaubhaftigkeitsmerkmalen insgesamt wenig übersichtlich und lässt sich kaum verallgemeinern. Vor diesem Hintergrund erscheint es ebenfalls problematisch, dass einzelne Studien zur Validität der CBCA gar keine Inter-Rater-Reliabilitäten berichten (z.B. Köhnken et al., 1995; Steller et al., 1992). Die Erfassung der ARJS erfolgt hingegen anhand standardisierter Ratingskalen. Dadurch lassen sich Variabilitäten in der Beurteilerübereinstimmung zwischen verschiedenen Studien auf die Beurteiler selbst oder das verwendete Stimulusmaterial zurückführen. Sporer und Burghardt (2004) verwendeten dasselbe Stimulusmaterial wie die vorliegende Studie und schulten ihre Beurteiler gleichermaßen intensiv. Infolgedessen überrascht es kaum, dass die Inter-Rater-Reliabilitäten beider Untersuchungen ähnlich gut ausfallen. Hingegen wurden von Sporer (1998) für einzelne Skalen etwas geringere Reliabilitäten berichtet (z.B. Details), was vermutlich durch die weniger umfangreiche Schulung der Beurteiler zu erklären ist.

Zusatzanalysen anhand der Spearman-Brown-Korrekturformel zeigten zudem, dass sich durch die Zusammenfassung mehrerer unabhängiger Beurteilungen die Inter-Rater-Reliabilitäten deutlich verbesserten. Bei den freien Berichten erzielten zehn, bei den Interviews elf der 13 ARJS-Skalen Reliabilitäten von $r \geq .70$. Zudem verfehlten die Skalen Komplikationen und ungewöhnliche Details bei beiden Aussageformen sowie die Skala Klarheit und Lebendigkeit bei den freien Berichten diesen Richtwert nur knapp. Lediglich für die Skala Realismus und logische Struktur war kein korrelativer Zusammenhang zwischen den Urteilen verschiedener Raterinnen nachweisbar. Allerdings ist dies auch inhaltlich vor dem Hintergrund des verwendeten Stimulusmaterials zu sehen. Da Aussagen von Erwachsenen analysiert wurden, überrascht es kaum, dass diese sehr realistisch und logisch kohärent waren. Untersuchungen zu Alltagsvorstellungen von Täuschungsindikatoren haben gezeigt, dass logische Konsistenz subjektiv als wichtiges Merkmal glaubhafter Aussagen erachtet wird

(Bond & DePaulo, 2006; vgl. Studie 2). Vermutlich achten Personen daher in besonderem Maße darauf, dass ihre Darstellungen logisch stimmig sind. Aufgrund eines Deckeneffektes und der fehlenden Variabilität hinsichtlich der Skala Realismus und logische Struktur war die Korrelation kein geeignetes Maß zur Abschätzung der Inter-Rater-Reliabilität. Durch Zusatzanalysen, die wie andere Autoren (z.B. Steller et al., 1992; Vrij, Kneller & Mann, 2000) prozentuale Übereinstimmungen als Maß der Inter-Rater-Reliabilität zugrunde legten, konnte jedoch eine hohe absolute Übereinstimmung in der Beurteilung der Skala Realismus und logische Struktur nachgewiesen werden.

Entscheidungen hinsichtlich der Glaubhaftigkeit einer Aussage werden in der Regel auf der Grundlage eines Sachverständigengutachtens getroffen. Von praktischem Interesse ist daher die Reliabilität einzelner Ratings. Diese wurde im Rahmen der vorliegenden Studie quantifiziert, indem die Korrelationen zwischen den Beurteilungen verschiedener Rater paarweise gemittelt wurden. Dabei ergaben sich durchaus angemessene Reliabilitäten für die Durchführung von forschungsorientierten Gruppenvergleichen. Allerdings stellt sich für individualdiagnostische Anwendungen die Frage, ob es möglicherweise günstiger wäre Entscheidungen auf der Grundlage mehrerer Gutachten zu treffen, oder zumindest ein Team zur Beurteilung von Aussagen einzusetzen, dessen gemittelte Ratings dann durch den verantwortlichen Sachverständigen benutzt werden. Die Reliabilität eines Urteils lässt sich systematisch erhöhen, wenn statt der Einzelratings die Urteile mehrerer Rater als Informationsgrundlage verwendet werden. Anhand der Spearman-Brown-Formel ist es möglich abzuschätzen, inwieweit sich die Reliabilität durch die Hinzunahme weiterer Rater verbessern lässt bzw. wie viele Rater notwendig sind, um einen akzeptablen Reliabilitätswert zu erzielen. Nach den vorliegenden Befunden dürfte sich bei drei bis vier Ratern eine für individualdiagnostische Zielsetzungen angemessene Inter-Rater-Reliabilität sicherstellen lassen. Allerdings erscheint dies insbesondere aufgrund ökonomischer Bedenken für die Praxis kaum umsetzbar.

Zusammenfassend ließen sich die ARJS reliabel auf die Analyse von Transkripten anwenden. Ebenso wurde demonstriert, dass sich die Inter-Rater-Reliabilität dadurch verbessern lässt, dass mehrere unabhängige Raterinnen für die Aussageanalyse eingesetzt werden. Die daraus resultierenden Reliabilitäten würden größtenteils den Anforderungen an individualdiagnostische Untersuchungen genügen (Wirtz & Caspar, 2002).

Validität der ARJS

Gemäß der theoretischen Entwicklung der ARJS wurde angenommen, dass die ARJS-Skalen eine höhere inhaltliche Qualität bei wahren als bei erfundenen Aussagen abbilden. Diese Annahme ließ sich für die meisten Skalen bestätigen. Die getrennte Analyse beider Aussageformen ergab für die freien Berichte bei vier, für die Interviews bei sieben Skalen signifikante Unterschiede in Abhängigkeit vom tatsächlichen Wahrheitsstatus. Da sich zwischen den freien Berichten und den Interviews jedoch keine bedeutsamen Unterschiede in der Validität der ARJS auf multivariater Ebene ergaben, wurden die beiden Aussageformen zusätzlich gemeinsam analysiert. Die resultierenden Effektstärken sind in Tabelle 3.13 den bisherigen Forschungsbefunden zur Validität der ARJS gegenübergestellt. Die Effektstärkemaße r entsprechen dabei der punktbiserialen Korrelation zwischen dem Wahrheitsstatus und der intervallskalierten Beurteilung der jeweiligen ARJS-Skala.

Bei der gemeinsamen Analyse der freien Berichte und der Interviews ergaben sich in der vorliegenden Studie erwartungsgemäße signifikante Unterschiede zwischen wahren und erfundenen Aussagen für neun Skalen. Dabei zeigten sich erwartungskonforme Befunde für ARJS-Skalen, die vor dem Hintergrund verschiedener Theorien konzipiert wurden.

Tabelle 3.13

Befunde bisheriger sowie der vorliegenden Studie zur Validität der ARJS (Effektstärkemaße r)

Skalen	Sporer (1998) N = 71	Sporer & Burghardt (2004) N = 184	Sporer & Walther (2006)[a] N = 72	Barnier et al. (2005)[b] N = 180	Vorliegende Studie N = 176	Mittleres ungewichtetes r[c]
Realismus und logische Struktur	.02	-.08	-.01	.07	.08	.02
Klarheit und Lebendigkeit	.05	.14	**.26**	.12	**.31**	.18
Details	.16	**.23**	**.40**	.14	**.32**	.25
Räumliche Details	-.08	.00	-.08	.10	.08	.00
Zeitliche Details	**.26**	**.17**	**.29**	.02	**.15**	.18
Sinneseindrücke	.12	.07	-.08	-.11	-.02	.00
Emotionen und Gefühle	**.24**	**.21**	**.38**	.13	**.25**	.24
Gedanken	.22	.04	**.26**	.10	**.15**	.16
Memorieren und Gedächtnis	.14	**.27**	**.45**	**.17**	**.18**	.25
Nonverbale und verbale Interaktionen	.08	.08	**.32**	.11	**.16**	.15
Komplikationen und ungewöhnliche Details	**.42**	**.19**	**.36**	.05	**.21**	.25
Fehler und sozial Unerwünschtes	**.39**	.09	-.02	.15	**.16**	.16
Persönliche Signifikanz	.20	.06	**.33**	-.04	-.11	.09

Anm. Signifikante Werte ($p < .05$) sind fett gedruckt. [a] = Die Effektstärken für die freien Berichte und die Interviews wurden zunächst Fisher Z-transformiert und dann gemittelt. [b] = Es wurde der Vergleich zwischen wahren und erfundenen Aussagen ausgewählt. [c] = Mittleres r basiert auf den zunächst Fisher Z-transformierten und dann über alle Studien gemittelten Effektstärken.

Im Folgenden werden die Befunde zu den einzelnen Skalen besprochen. Die dabei aufgeführten Verweise auf deren theoretischen Entwicklungshintergrund basieren auf der Zusammenfassung von Sporer (2004) sowie auf Vortragsmanuskripten (Sporer, 1998; Sporer & Burghardt, 2004; Sporer, Samweber & Stucke, 2002) und der persönlichen Kommunikation mit dem Autor der ARJS.

Die Validität der ARJS-Skalen Klarheit und Lebendigkeit, zeitliche Details, Emotionen und Gefühle sowie Memorieren und Gedächtnis verweist darauf, dass der RÜ-Ansatz und andere gedächtnispsychologische Theorien auf die Entdeckung von Täuschung anwendbar sind. Der RÜ-Ansatz postuliert konkrete Merkmalsunterschiede zwischen vorgestellten und erlebnisbasierten Erinnerungen. Diese wurden traditionell durch Selbstbeschreibungen von Erinnerungsqualitäten überprüft und später auf die Fremdbeurteilung von Erlebnisberichten übertragen (z.B. Alonso-Quecuty, 1992; Sporer & Küpper, 1995; Sporer & Sharman, 2006; vgl. auch Masip et al., 2005). Signifikante Lebensereignisse, wie die in der vorliegenden Studie verwendeten, sind vermutlich persönlich bedeutsam für die Konzeption des Selbst und autobiographischer Strukturen. Der gedächtnispsychologischen Literatur wiederum ist zu entnehmen, wie das autobiographische Gedächtnis organisiert ist, welche Faktoren die Gedächtnisleistung verbessern und durch welche Merkmale sich autobiographische Erinnerungen charakterisieren lassen. Im Rahmen der ARJS wird angenommen, dass diese Merkmale eher bei wahren als bei erfundenen Aussagen vorzufinden sind. Diese Argumentation setzt jedoch voraus, dass Personen nicht willens oder nicht fähig sind, entsprechende Merkmale zu erfinden. Demnach erscheint es sinnvoll zu berücksichtigen, welche Merkmale Personen mit wahren bzw. erfundenen Aussagen assoziieren. Wenn Merkmale subjektiv als Lügenindikatoren aufgefasst werden, so ist nicht zu befürchten, dass Lügner sie gezielt integrieren und deren Validität infolgedessen vermindert wird. Leider liegt für die ARJS-Merkmale bislang nur eine Studie vor, in

der subjektive Annahmen zu Täuschungskorrelaten erfasst wurden (vgl. Studie 2). Soweit es aufschlussreich erscheint, werden deren Befunde bei den nachfolgenden Ausführungen berichtet.

Für die ARJS-Skala Klarheit und Lebendigkeit ließ sich ein starker Effekt nachweisen. Die Bedeutsamkeit der ihr zugehörigen Merkmale wurde beispielsweise in der Literatur zum autobiographischen Gedächtnis herausgestellt. So argumentierte Brewer (1986, 1996), dass Erinnerungen an selbst-erlebte Ereignisse oftmals vom Eindruck des Wiedererlebens begleitet sind. Dabei betonte der Autor besonders die Bedeutsamkeit visueller Eindrücke, die auch bei der ARJS-Skala Klarheit und Lebendigkeit Berücksichtigung finden. Brewer (1988, Experiment 2) konnte empirisch nachweisen, dass visuelle Eindrücke bei richtigen Erinnerungen an autobiographische Ereignisse stärker ausgeprägt sind als bei falschen. Zudem stellten Studien zum RÜ-Ansatz wiederholt fest, dass selbst-erlebte im Vergleich zu vorgestellten Ereignissen als lebendiger und klarer beschrieben wurden (Johnson & Raye, 1981; Larsen, 1998; McGinnis & Roberts, 1996; Suengas & Johnson, 1988). Der für die ARJS ermittelte Befund einer höheren Klarheit und Lebendigkeit wahrer im Vergleich zu erfundenen Aussagen zeigt, dass diese Befunde auf Fremdbeurteilungen von Aussagen generalisierbar sind.

Zudem erscheint eine getrennte Erfassung zeitlicher und räumlicher Details aufgrund der vorliegenden Untersuchungsbefunde sinnvoll. Wahre Aussagen erzielten erwartungsgemäß höhere Beurteilungen hinsichtlich der ARJS-Skala zeitliche Details als erfundene. Deren Validität lässt sich erneut über gedächtnispsychologische Theorien begründen. Verschiedene Modelle zum autobiographischen Gedächtnis gehen davon aus, dass neben thematischen Bezügen auch zeitlichen Informationen strukturierende Bedeutung zukommt (vgl. Conway & Pleydell-Pearce, 2000; Pohl, 2007). Diese Annahme wurde beispielsweise von Anderson und Conway (1993) empirisch untermauert. Die Autoren überprüften in mehreren Experimenten, wie viele Details zu spezifischen

Ereignissen, in welcher Reihenfolge und mit welcher Geschwindigkeit abgerufen wurden. Dabei zeigte sich, dass Probanden entweder zunächst distinkte und persönlich bedeutsame Details erinnerten, oder dass der Abruf chronologisch vom ersten bis zum letzten Detail organisiert war. Entsprechend ist davon auszugehen, dass zeitliche Informationen ein wichtiges Merkmal erlebnisbasierter Gedächtnisinhalte darstellen. Dadurch lässt sich erklären, dass sie auch in der vorliegenden Untersuchung eher bei wahren als bei erfundenen Aussagen vorzufinden waren.

Nach dem RÜ-Ansatz sollten sich auch für räumliche Details Unterschiede zwischen wahren und erfundenen Aussagen zeigen (Johnson & Raye, 1981; Johnson, Hashtroudi & Lindsay, 1993; vgl. auch Masip et al., 2005; Vrij, Edward et al., 2000, Vrij et al., 2001b). Im Rahmen der vorliegenden Studie war dies jedoch nicht festzustellen. Auch andere Studien konnten die Validität der ARJS-Skala räumliche Details nicht belegen (Sporer, 1998; Sporer & Burghardt, 2004; Sporer & Walther, 2006). Aufgrund der zeitlichen Organisationsstruktur des autobiographischen Gedächtnis wäre es durchaus denkbar, dass zeitliche Details eine validere Unterscheidung wahrer und erfundener Aussagen erlauben als räumliche. Zudem stellt sich die Frage, ob sich wahre und erfundene Aussagen besser unterscheiden lassen, wenn beide Merkmale gemeinsam auftreten, als wenn nur Informationen zu einem Bereich vorliegen (Sporer, 1997a). Dies lässt sich bei gemeinsamer Betrachtung zeitlicher und räumlicher Details nicht feststellen. Daher erscheint es problematisch, dass die CBCA nicht zwischen diesen beiden Aussagemerkmalen differenziert und auch in einzelnen RÜ-Studien beide Aspekte gemeinsam betrachtet wurden (z.B. Alonso-Quecuty, 1992). Infolgedessen ist es als wesentlicher Vorteil der ARJS aufzufassen, dass zeitliche und räumliche Merkmale getrennt voneinander erfasst werden.

Des Weiteren fanden sich in der vorliegenden Untersuchung bei wahren Aussagen erwartungsgemäß mehr Verweise auf Emotionen und Gefühle als bei erfundenen. Dies lässt sich beispielsweise anhand des Gedächtnismodells von

Johnson (1985) erklären, das eine Enkodierung affektiver Informationen im Gedächtnis postuliert. Infolgedessen sind emotionale Informationen eher bei wahren als bei erfundenen Aussagen zu erwarten. Diese Annahme wurde durch einige Untersuchungen zum RÜ-Ansatz empirisch gestützt. Bei selbst-erlebten Ereignissen wurden mehr affektive Informationen berichtet als bei vorgestellten bzw. erfundenen Ereignissen (z.B. Larsson & Granhag, 2005; Sporer, 1997a). Andere Studien zum RÜ-Ansatz differenzierten allerdings nicht zwischen Gedanken und Gefühlen (z.B. Hernandez-Fernaud & Alonso Quecuty, 1997; McGinnis & Roberts, 1996). In der vorliegenden Studie zeigte sich für die ARJS-Skala Emotionen und Gefühle ein deutlicher Effekt mittlerer Größenordnung, während der Effekt für die Skala Gedanken geringer war. Demnach erscheint es sinnvoll diese Aussagemerkmale eigenständig zu erfassen.

Auch die ARJS-Skala Memorieren und Gedächtnis erzielte höhere Beurteilungen bei wahren als bei erfundenen Aussagen. Diese Skala erfasst unter anderem, ob wiederholt über das Ereignis nachgedacht und gesprochen wurde. In der gedächtnispsychologischen Literatur wurde darauf hingewiesen, dass durch solche Aktivitäten die Gedächtnisspur gefestigt und der Informationsabruf erleichtert wird (z.B. Suengas & Johnson, 1988). Zudem wird bei der ARJS-Skala Memorieren und Gedächtnis berücksichtigt, ob Ereignisse vor und nach dem Hauptgeschehen sowie unterstützende Erinnerungen geschildert werden. Solche Aspekte verweisen darauf, dass Gedächtnisinhalte in einem assoziativen Netzwerk verankert sind (vgl. Anderson, 1983; Bower & Hilgard, 1983). Entsprechend konnte gezeigt werden, dass konzeptionell ähnliche RÜ-Merkmale bei erlebnisbasierten Ereignissen stärker ausgeprägt waren als bei vorgestellten (Johnson et al., 1988, Exp. 1). Solche Unterschiede in der Selbstbeschreibung von Erinnerungsqualitäten scheinen sich nach den vorliegenden Befunden auch in der Fremdbeurteilung erlebnisbasierter und vorgestellter Aussagen abzubilden.

Wahre Aussagen erzielten zudem höhere Bewertungen hinsichtlich der ARJS-Skala Komplikationen und ungewöhnliche Details als erfundene. Die

Validität dieser Skala lässt sich erneut gedächtnispsychologisch begründen. In der sozialen Kognitionsforschung zeigte sich, dass schema-relevante Informationen besser erinnert werden als schema-irrelevante (Hastie, 1981). Bei schema-relevanten Informationen ist zudem zu beachten, ob sie dem Schemawissen entsprechen (schemakonsistente Informationen) oder dieses verletzen (schemainkonsistente Informationen; vgl. Fiske & Taylor, 1991). Wenn beispielsweise das Schema eines Priesters aktiviert wurde, wäre es als schemakonsistente Information zu werten, dass dieser eine Kirche besucht. Hingegen wäre es eine schemainkonsistente Information, den Priester in einer verrauchten Kneipe beim Bier anzutreffen. Dieses Beispiel wurde im ARJS-Manual (Sporer, 1996/1998/2004) zur Erläuterung des Aussagemerkmals ungewöhnliche Details angeführt.

Zwei Metaanalysen, die sich vor allem in ihren Ausschlusskriterien für die Literaturauswahl sowie den Moderatoranalysen unterschieden, untersuchten die Erinnerung an schemarelevante Informationen (Stangor & McMillan, 1992; Rojahn & Pettigrew, 1992). Insgesamt stellten beide Metaanalysen einen Erinnerungsvorteil für schemainkonsistente im Vergleich zu schemakonsistenten Informationen fest, wenn die Gedächtnistests einen freien Abruf von Informationen erforderten. Dies wurde darauf zurückgeführt, dass schemainkonsistente Informationen erklärungsbedürftig sind und daher mit vielfältigen anderen Informationen verknüpft werden (z.B. Srull & Wyer, 1989). Die unter der Skala Komplikationen und ungewöhnliche Details zusammengefassten Aussagemerkmale sind als schemainkonsistent zu werten. Falls selbst-erlebte Ereignisse mit Komplikationen und ungewöhnlichen Details einhergingen, sollten diese als schemainkonsistente Informationen auch erinnert werden. Zudem gibt es Hinweise darauf, dass Personen Komplikationen und ungewöhnliche Details subjektiv mit erfundenen Aussagen assoziieren (z.B. Akehurst, Köhnken, Vrij & Bull, 1996; Taylor & Vrij, 2000). Dies konnte auch speziell für die ARJS-Merkmale nachgewiesen werden (vgl. Studie 2). Daher ist anzunehmen, dass Lügner

entsprechende Informationen nicht erfinden würden. Der im Rahmen der vorliegenden Studie ermittelte Effekt mittlerer Größenordnung für die ARJS-Skala Komplikationen und ungewöhnliche Details unterstützt diese Argumentation.

Die bisher dargestellten Befunde zu den ARJS stehen im Einklang mit den theoretischen Annahmen des RÜ-Ansatzes. Dies gilt jedoch nicht für die ARJS-Skala Gedanken. Die ARJS-Skala Gedanken war bei wahren Aussagen erwartungsgemäß höher ausgeprägt als bei erfundenen. Dieser Befund stützt die den ARJS zugrunde liegende Annahme, dass die entsprechenden Merkmale als Wahrheitsindikatoren aufzufassen sind. Nach dem RÜ-Ansatz wäre Gegensätzliches zu erwarten gewesen (Johnson, 2006; Johnson & Raye, 1981). Bezogen auf die Fremdbeurteilung von Aussagen ließ sich diese RÜ-Annahme allerdings kaum empirisch unterstützen. Bisherige Studien zur Entdeckung von Täuschung fanden für das RÜ-Merkmal kognitive Operationen nur selten Unterschiede zwischen wahren und erfundenen Aussagen (vgl. Masip et al., 2005). In einzelnen Untersuchungen waren kognitive Operationen sogar als Wahrheitsindikator aufzufassen (Sporer & Küpper, 2004; Vrij, Edward et al., 2000). Daher argumentierte Sporer (2004), dass es notwendig ist, zwischen verschiedenen Arten kognitiver Operationen zu differenzieren. Insbesondere stützende Erinnerungen und Memorierungsprozesse festigen seines Erachtens die Erinnerung an selbst-erlebte Ereignisse und sind daher eher bei wahrheitsgemäßen Aussagen zu erwarten. Die unter der ARJS-Skala Gedanken zusammengefassten Merkmale sind als Hinweise auf stützende Erinnerungen (vgl. Johnson, 1985) und damit als Wahrheitsindikatoren aufzufassen. Diese Annahme wird durch die vorliegenden Untersuchungsbefunde unterstützt.

Der RÜ-Ansatz zielte ursprünglich darauf ab zu erklären, wie Personen eigene intern und extern generierte Erinnerungen voneinander unterscheiden. Für eine Übertragung der Vorhersagen auf die interpersonelle Ebene scheinen Modifikationen notwendig zu sein. Nach den vorliegenden Untersuchungsbefunden können sozialpsychologische Überlegungen zur

Entdeckung von Täuschung den RÜ-Ansatz offenbar sinnvoll ergänzen. So zeigten sich erwartungsgemäße Unterschiede für die ARJS-Skala Fehler und sozial Unerwünschtes. Diese bezieht sich auf Korrekturen oder Präzisierungen sowie auf einen Mangel an sozialer Erwünschtheit. Des Weiteren werden entgegen der Vorhersagen des RÜ-Ansatzes (Johnson & Raye, 1981) jedoch im Einklang mit den CBCA, Erinnerungslücken als Wahrheitsindikatoren aufgefasst. Auch widerspricht es dem RÜ-Ansatz, dass Verweise auf die Sicherheit der eigenen Angaben eher bei erfundenen als bei wahren Aussagen erwartet werden. Die Validität der Skala Fehler und sozial Unerwünschtes führt Sporer (1996) auf sozialpsychologische Theorien zurück. So argumentiert er vor dem Hintergrund des Impression-Management-Ansatzes, dass die dazugehörigen Merkmale stereotypen Täuschungsindikatoren entsprechen und daher von Lügnern gezielt vermieden werden (vgl. auch DePaulo et al., 2003). So verletzt beispielsweise das Zugeben von Erinnerungslücken das Kommunikationsmaxim einer vollständigen Darstellung (Grice, 1975). Entsprechend zeigten einige Untersuchungen zu subjektiven Täuschungsindikatoren, dass Laien die entsprechenden Merkmale mit erfundenen Aussagen assoziieren (z.B. Akehurst et al., 1996; Taylor & Vrij, 2000; Vrij, Akehurst & Knight, 2006). Allerdings liegen auch Forschungsbefunde vor, die diese Argumentation von Sporer (1996) nicht unterstützen (z.B. Lakhani & Taylor, 2003; Vrij et al., 2001a; vgl. auch Studie 2). Dennoch erzielte die Skala Fehler und sozial Unerwünschtes im Rahmen der vorliegenden Studie eine gute Validität. Daher erscheint es sinnvoll, sozialpsychologische Überlegungen bei der Konzeption von Glaubhaftigkeitsmerkmalen zu berücksichtigen.

Schließlich ergaben sich erwartungsgemäße Unterschiede zwischen wahren und erfundenen Aussagen für ARJS-Skalen, die den CBCA-Merkmalen konzeptionell ähnlich sind. Dies verweist darauf, dass auch aufgrund von praktischem Erfahrungswissen konzipierte Merkmale wahre und erfundene Aussagen voneinander unterscheiden können. So wurde ein starker Effekt für die ARJS-Skala Details festgestellt, wobei Aussagen zu selbst-erlebten Ereignissen

erwartungsgemäß detaillierter waren als zu erfundenen. Diese Skala weist konzeptionelle Ähnlichkeiten zu den beiden CBCA-Merkmalen quantitativer Detailreichtum und Schilderung nebensächlicher Einzelheiten auf. Vrij (2005) schlussfolgerte aus seiner qualitativen Literaturübersicht, dass das CBCA-Merkmal des quantitativen Detailreichtums bislang die meiste empirische Unterstützung erfahren hat. Allerdings finden sich in seiner Literaturübersicht auch Untersuchungen, deren Befunde mit Vorsicht zu interpretieren sind. Beispielsweise kann man durchaus von einem positiven Zusammenhang zwischen dem Umfang von Aussagen und ihrem Detaillierungsgrad ausgehen. Das CBCA-Kriterium des quantitativen Detailreichtums mit der Anzahl der Wörter gleichzusetzen (Santilla et al., 2000), erscheint jedoch problematisch. Dennoch steht der vorliegende Befund einer guten Unterscheidbarkeit wahrer und erfundener Aussagen hinsichtlich der ARJS-Skala Details im Einklang mit der Schlussfolgerung von Vrij (2005). Zudem hat sich die sehr präzise operationalisierte ARJS-Skala nonverbale und verbale Interaktionen als valide erwiesen. Sie erfasst ähnliche Aspekte wie die beiden CBCA-Merkmale Interaktionsschilderung und Wiedergabe von Gesprächen, deren Validität ebenfalls wiederholt nachgewiesen wurde (vgl. Vrij, 2005). Da die beiden CBCA-Merkmale jedoch meist zusammen auftreten sind sie bei den ARJS zu einer Skala zusammengefasst (Sporer, 2004).

Während die bisher dargestellten Befunde einen überzeugenden Validitätsbeleg für die ARJS darstellen, ließen sich für drei Skalen keine erwartungsgemäßen Unterschiede zwischen wahren und erfundenen Aussagen nachweisen. Hinsichtlich der Skala persönliche Signifikanz ergab sich ein erwartungskonträrer, jedoch nicht signifikanter Mittelwertsunterschied. Erfundene Aussagen erzielten tendenziell höhere Beurteilungen als wahre. Einen möglichen Ansatzpunkt zur Erklärung dieses Befundes liefert eine Untersuchung von Destun und Kuiper (1999). Die Autoren instruierten Erwachsene, sich Ereignisse mit negativer und positiver Valenz sowohl vorzustellen und als auch wahrheitsgemäß

daran zu erinnern. Bei vorgestellten Ereignissen wurden vor allem Emotionen berichtet, die mit der vorgegebenen Valenz im Einklang standen. Wenn sich die Probanden negative Ereignisse vorstellten, berichteten sie also nur negative Emotionen, bzw. bei positiven Ereignissen nur positive Emotionen. Hingegen fanden sich bei erlebnisbasierten Erinnerungen auch Verweise auf Emotionen, die im Widerspruch zur vorgegebenen Valenz standen. So beinhalteten Erinnerungen an erlebnisbasierte positive Ereignisse auch negative Gefühle, bzw. Erinnerungen an negative Ereignisse auch positive Gefühle. Für die vorliegende Studie ließe sich spekulieren, dass sich wahre und erfundene Aussagen analog zu den Befunden von Destun und Kuiper (1999) in der Darstellung der persönlichen Signifikanz unterschieden. Möglicherweise stellten Personen, die Erfundenes berichteten, die Ereignisse als äußerst bedeutsam dar. Hingegen wäre es denkbar, dass die Bedeutsamkeit erlebnisbasierter Ereignisse als vergleichsweise komplex dargestellt wird. So können persönliche Erlebnisse zwar zum Zeitpunkt des Geschehens als äußerst bedeutsam empfunden werden, sich im Nachhinein jedoch als weniger bedeutsam erweisen. Möglicherweise wurden hinsichtlich der ARJS-Skala persönliche Signifikanz höhere Beurteilungen vergeben, wenn Personen die Bedeutsamkeit erfundener Ereignisse klar herausstellten, als wenn sie eine differenziertere Sichtweise erfahrungsbasierter Ereignisse zum Ausdruck brachten. Trotz dieser Überlegungen bleibt festzuhalten, dass der Mittelwertsunterschied für die ARJS-Skala persönliche Signifikanz keine statistische Signifikanz erzielte. Dies galt sowohl für die getrennte Analyse beider Aussageformen als auch für deren gemeinsame Betrachtung. Allerdings fanden andere Studien der theoretischen Erwartung entsprechend tendenziell bzw. signifikant mehr Hinweise auf die persönliche Signifikanz bei wahren im Vergleich zu erfundenen Aussagen (Sporer, 1998; Sporer & Walther, 2006).

Schließlich bildeten die Skalen Realismus und logische Struktur und Sinneseindrücke keinerlei Unterschiede zwischen wahren und erfundenen

Aussagen ab. Allerdings ist dies, wie bereits im Zusammenhang mit den Befunden zur Reliabilität diskutiert, auf Decken- und Bodeneffekte zurückzuführen.

Um einen Überblick hinsichtlich des derzeitigen Forschungsstandes zur Validität der ARJS zu gewinnen, wurden in Tabelle 3.13 die vorliegenden Untersuchungsbefunde mit denen der bisherigen Studien integriert. Dazu wurde für jede Skala eine ungewichtete mittlere Effektstärke über alle Studien hinweg berechnet. Die in den einzelnen Studien berichteten Effektstärken wurden dazu Fisher's -Z transformiert, danach gemittelt und schließlich in r zurücktransformiert. Die stärksten Mittelwertsunterschiede zwischen wahren und erfundenen Aussagen wurden für die Skalen Details, Memorieren und Gedächtnis, Emotionen und Gefühle, Komplikationen und ungewöhnliche Details sowie zeitliche Details nachgewiesen. Doch auch die Skalen Klarheit und Lebendigkeit, nonverbale und verbale Interaktionen, Gedanken sowie Fehler und sozial Unerwünschtes erwiesen sich wiederholt als valide. Hingegen ließen sich nur in jeweils einer Untersuchung geringe Effekte des Wahrheitsstatus auf die beiden Skalen Sinneseindrücke (Sporer, 1998) und räumliche Details (Barnier et al., 2005) feststellen.

Für die Skala persönliche Signifikanz zeigten sich in der vorliegenden Untersuchung tendenziell erwartungskonträre Befunde. Im Gegensatz dazu verweist jedoch die Untersuchung von Sporer und Walther (2006) darauf, dass diese Skala erwartungsgemäß zwischen wahren und erfundenen Aussagen unterscheidet. Auch Sporer (1998) fand hinsichtlich der persönlichen Signifikanz höhere Beurteilungen für wahre als für erfundene Aussagen, jedoch ließ sich dieser Mittelwertsunterschied nicht statistisch absichern. Die Validität der Skala Realismus und logische Struktur ließ sich durch die in Tabelle 3.13 aufgeführten Studien nicht belegen. Allerdings wurde bei der tabellarischen Übersicht auf eine ARJS-Studie von Sporer, Samweber et al. (2000) verzichtet, weil dabei keine vollständige ARJS-Analyse durchgeführt wurde. Vielmehr wurden die Aussagen von Laien beurteilt, die nur kurz über neun ARJS-Kriterien schriftlich informiert

worden waren. Diese Beurteilungen bildeten jedoch erwartungsgemäße Unterschiede zwischen wahren und erfundenen Aussagen für das Merkmal Realismus und logische Struktur ab. Vermutlich lassen sich die Schwierigkeiten im Validitätsnachweis für die ARJS-Skala Realismus und logische Struktur darauf zurückführen, dass bislang nur Aussagen von Erwachsenen analysiert wurden (Barnier et al., 2005; Sporer, 1998; Sporer & Burghardt, 2004; Sporer & Walther, 2006). Möglicherweise erlauben es die entsprechenden Aussagemerkmale jedoch wahre und erfundene Aussagen von Kindern zu unterscheiden (Sporer, 2004).

Zusammenfassend ergaben sich in der vorliegenden Untersuchung für neun der 13 ARJS-Skalen erwartungsgemäße Mittelwertsunterschiede zwischen wahren und erfundenen Aussagen. Dies stellt einen überzeugenden Validitätsnachweis für die ARJS dar. In der einschlägigen Literatur zum Thema Täuschung wurden häufig nur schwache Effekte erzielt. Beispielsweise berichteten DePaulo et al. (2003), dass der Median der Effektstärke von 88 untersuchten Täuschungsindikatoren bei d = 0.10 lag, was einer punktbiserialen Korrelation von r = .05 entspricht (Rice & Harris, 2005; Rosenthal, 1994). Im Vergleich dazu erscheinen die in der vorliegenden Untersuchung ermittelten Effektstärken für inhaltliche Aussagemerkmale durchaus beindruckend (M(r) = .15). Ebenso verweist die überzufällig hohe Klassifikationsrate mittels multipler Diskriminanzanalysen auf die Nützlichkeit der ARJS, wahre und erfundene Aussagen zu unterscheiden.

Effekte der Vorbereitungszeit

Bei der Untersuchung inhaltlicher Glaubhaftigkeitsmerkmale erscheint es zudem erforderlich potenzielle Moderatoren in Betracht zu ziehen. So ist beispielsweise anzunehmen, dass sich Personen vorbereiten, wenn sie juristisch relevante Aussagen treffen. Daher wurde die Vorbereitungszeit als Between-Subjects-Faktor in der vorliegenden Untersuchung experimentell variiert (vgl. Alonso-Quecuty, 1992; Sporer, 1998).

Den Stimuluspersonen wurden unmittelbar vor den freien Berichten entweder zwei oder 15 Minuten zur Vorbereitung ihrer Aussage eingeräumt. Es wurde erwartet, dass sich die inhaltliche Aussagequalität durch Vorbereitung verbessern lässt. Zudem sollte geklärt werden, ob sich die Gelegenheit zur Vorbereitung validitätsmindernd auf die ARJS auswirkt.

Erwartungsgemäß war eine höhere Aussagequalität vorbereiteter freier Berichte im Vergleich zu unvorbereiteten festzustellen. Dies war unter anderem darauf zurückzuführen, dass die Berichte mit Details angereichert wurden. So ergab sich der stärkste Effekt für die Skala Details, gefolgt von den Skalen Memorieren und Gedächtnis, nonverbale und verbale Interaktionen und zeitliche Details. Diese Art von Details wahrheitsgemäß abzurufen oder frei zu erfinden wird offenbar durch die Gelegenheit zur Vorbereitung gefördert.

Eine Wechselwirkung zwischen der Vorbereitungszeit und dem Wahrheitsstatus auf die ARJS-Beurteilungen wurde jedoch nicht festgestellt. Die ARJS-Aussageanalyse erlaubte es demnach, unabhängig von der Gelegenheit zur Vorbereitung wahre und erfundene Aussagen zu unterscheiden (vgl. Sporer & Burghardt, 2004). Anscheinend nutzten alle Personen die Vorbereitungszeit, um die Qualität ihrer Aussage zu erhöhen. Es scheint jedoch grundsätzlich leichter zu sein, sich an erlebnisbasierte Details zu erinnern, als neue frei zu erfinden. Dies würde erklären, dass die ARJS auch bei vorbereiteten Aussagen erwartungsgemäße Unterschiede in Abhängigkeit vom Wahrheitsstatus abbildeten.

Allerdings wurden in der Literatur zu inhaltlichen Aussagemerkmalen wiederholt Wechselwirkungen zwischen dem Wahrheitsstatus und der Gelegenheit zur Vorbereitung auf die Aussagequalität berichtet. So konnte Alonso-Quecuty (1992) die Validität der RÜ-Kriterien sensorische und kontextuelle Informationen nur für unvorbereitete Aussagen nachweisen. Eine 10-minütige Vorbereitungszeit reichte aus, um entsprechende Unterschiede zwischen wahren und erfundenen Aussagen zu nivellieren. Allerdings erscheint die ökologische

Validität dieser Studie fraglich. Im Rahmen eines Within-Subjects-Designs gaben alle Probanden sowohl wahrheitsgemäße als auch falsche Aussagen über einen Filmausschnitt ab, der ihnen kurz zuvor gezeigt wurde. Die wahren Aussagen stellten eine Zusammenfassung der Inhalte dar, wohingegen für die falschen Aussagen nur ein Detail des Filmausschnitts abzuändern war. Falschaussagen dieser Art zu formulieren, sollte nicht besonders anspruchsvoll sein. Möglicherweise ist es darauf zurückzuführen, dass die RÜ-Kriterien nach einer kurzen Vorbereitung nicht mehr valide waren.

Im Gegensatz dazu verglich Sporer (1998) im Rahmen eines Between-Subjects-Designs Aussagen zu einem Ereignis, das die Probanden entweder tatsächlich selbst erlebt hatten oder nur vorgaben erlebt zu haben. Sporer (1998) berichtete vor allem für unvorbereitete Aussagen signifikante Effekte. Bei unvorbereiteten Aussagen zeigten sich für sieben ARJS-Skalen signifikante Unterschiede zwischen wahren und erfundenen Aussagen. Vorbereitete Teilnehmer wurden bereits am Vorabend über das zu schildernde Ereignis informiert. Dadurch konnten sie sich vergleichsweise gründlich auf ihre Aussage vorbereiten und möglicherweise auch andere Lehrgangsteilnehmer zu dem darzustellenden Ereignis befragen. Derart vorbereitete wahre und erfundene Aussagen unterschieden sich nur noch bezüglich der Skala ungewöhnliche Details erwartungsgemäß. Im Gegensatz dazu verweisen die Befunde von Sporer und Küpper (1995) sowie von Sporer (1997a) auf eine bessere Validität inhaltlicher Aussagemerkmale bei vorbereiteten Berichten.

Insgesamt sind die bisherigen Untersuchungen zum Einfluss der Gelegenheit zur Vorbereitung widersprüchlich. Dies könnte unter anderem auf Unterschiede in der Dauer der Vorbereitung zurückzuführen sein. Zudem war der Stichprobenumfang in den aufgeführten Untersuchungen recht gering. Lediglich Sporer und Küpper (1995) führten eine umfangreichere Untersuchung durch und analysierten 200 Aussagen. Die Autoren wiesen jedoch selbst darauf hin, dass die einwöchige Gelegenheit zur Vorbereitung von den Probanden möglicherweise

nicht genutzt wurde. Dies sollte im Rahmen der vorliegenden Untersuchung vermieden werden, indem den Probanden nur 15 Minuten Gelegenheit gegeben wurde, ihre Aussage vorzubereiten.

In der vorliegenden Untersuchung gingen die Unterschiede in der Aussagequalität zudem mit Unterschieden in der Anzahl der Wörter einher. Vorbereitete Aussagen waren umfangreicher als unvorbereitete, auch wenn erneut keine Wechselwirkung zwischen der Vorbereitungszeit und dem Wahrheitsstatus vorlag. Im Gegensatz dazu verweisen metaanalytische Befunde darauf, dass es durchaus zu solchen Wechselwirkungseffekten kommen kann (Zuckerman & Driver, 1985; Sporer & Schwandt, 2006). Jedoch wurden dabei nur Effektstärken geringer Größenordnung ermittelt und die meisten Studien verwendeten relativ kurze Berichte.

Für die Interviews ergaben sich im Gegensatz zu den freien Berichten keine Unterschiede zwischen vorbereiteten und unvorbereiteten Aussagen. Dies resultierte in einer signifikanten Wechselwirkung zwischen der Aussageform und der Gelegenheit zur Vorbereitung. Allerdings ist diese Wechselwirkung vermutlich nicht auf die Art der Aussage zurückzuführen, sondern vielmehr auf die Tatsache, dass die Interviews eine Woche nach der experimentellen Variation der Vorbereitungszeit durchgeführt wurden. Offenbar waren die Effekte der Planung und Vorbereitung nicht nachhaltig genug, um noch nach einer Woche Qualitätsunterschiede zwischen ursprünglich vorbereiteten und unvorbereiteten Aussagen zu bewirken (vgl. Sporer & Burghardt, 2004).

Effekte der Aussageform

In der vorliegenden Untersuchung waren die Befunde zur Validität der ARJS über beide Aussageformen hinweg generalisierbar. Dies spiegelte sich auch in einer vergleichbaren Klassifikationsgüte der ARJS für die freien Berichte und die Interviews wider. Allerdings wäre es vorschnell anzunehmen, die ARJS seien gegenüber verschiedenen Befragungsformen robust. So zeigte sich bei den univariaten Analysen eine signifikante Wechselwirkung der Aussageform und des

Wahrheitsstatus auf die Skala Fehler und sozial Unerwünschtes. Für die freien Berichte ergaben sich hinsichtlich dieser Skala keinerlei Unterschiede. Bei den Interviews hingegen erzielten wahre Aussagen erwartungsgemäß höhere Beurteilungen als erfundene.

Die Validität der Skala Fehler und sozial Unerwünschtes wird auf Selbstdarstellungsbemühungen zurückgeführt, wobei der empirische Forschungsstand hinsichtlich dieser Argumentation widersprüchlich ist (z.B. Akehurst et al., 1996; Taylor & Vrij, 2000; Vrij et al., 2006; Vrij et al., 2001a; vgl. auch Studie 2). Hohe Bewertungen indizieren, dass Korrekturen oder Präzisierungen vorgenommen wurden, die Richtigkeit der eigenen Angaben in Frage gestellt wurde, Erinnerungslücken und persönliche Schwächen zugegeben wurden. Es wird argumentiert, dass diese Merkmale einen negativen Eindruck hinterlassen könnten und daher von Lügnern vermieden werden. Wahr aussagende Personen sollten vergleichsweise weniger um eine positive Selbstdarstellung bemüht sein. Infolgedessen sollten wahre Aussagen höhere Beurteilungen hinsichtlich dieser Skala erzielen als erfundene.

Die Interaktion mit einem Gesprächspartner könnte sich jedoch auf das Ausmaß an Selbstdarstellungsbemühungen und damit auch auf die Validität dieser Skala auswirken. So ist anzunehmen, dass interaktive Situationen einen besonders starken Aufforderungscharakter hinsichtlich sozialer Konventionen aufweisen. Möglicherweise kommt es nur dann zu verstärkten Selbstdarstellungsbemühungen von lügenden im Vergleich zu wahr aussagenden Personen, wenn Anforderungen an sozial erwünschtes Verhalten salient sind. Im Gegensatz zu freien Berichten kommt es im Rahmen von Interviews zwangsläufig zu einer Interaktion der aussagenden und befragenden Person. Die Form der Befragung könnte daher den Zusammenhang zwischen der ARJS-Skala und dem Wahrheitsstatus moderieren.

Bisherige Studien zu inhaltlichen Aussagemerkmalen haben vor allem deren Validität für spezifische Interviewtechniken überprüft. Dabei hat

insbesondere das Kognitive Interview nach Geiselman et al. (1986; vgl. auch Fisher & Geiselman, 1992) Forschungsaufmerksamkeit erfahren. So wurde überprüft, ob die Validität von CBCA- (Colwell et al., 2002; Köhnken et al., 1995; Steller & Wellershaus, 1996; vgl. auch Zaparniuk, Yuille & Taylor, 1995, zur Validität der CBCA beim schrittweisen Interview) und RÜ-Merkmalen (Hernandez-Fernaud & Alonso-Quecuty, 1997; Larsson & Granhag, 2005) durch das kognitive Interview beeinflusst wird. Insgesamt legen die Befunde nahe, dass die spezifischen Befragungstechniken des Kognitiven Interviews die Aussagequalität beeinflussen können. Hinsichtlich der Auswirkungen des Kognitiven Interviews auf die Validität inhaltlicher Aussagemerkmale wurden widersprüchliche Schlussfolgerungen gezogen (z.B. Köhnken et al., 1995; Steller & Wellershaus, 1996).

Nach einer Untersuchung von Anson et al. (1993) finden sich jedoch in der Praxis sehr unterschiedliche Interviewtechniken. Daher wurden im Rahmen der vorliegenden Studie lediglich freie Berichte und semi-strukturierte Interviews vergleichend betrachtet. Die Interviewerin wurde instruiert, auf suggestive Fragen zu verzichten und weiterführende Informationen hinsichtlich der Ereignisse zu erfragen. Eine nachträgliche Analyse verwies darauf, dass sie den Probandinnnen in fast allen Interviews durch offene Fragen Gelegenheit einräumte, weitere Informationen zu ergänzen. Zudem stellte die Interviewerin oftmals spezifische Fragen zu Zeit und Ort des Geschehens sowie zu beteiligten Personen. Die übrigen inhaltlichen Fragekategorien machten jeweils weniger als 5% der Äußerungen der Interviewerin aus. Daran wird deutlich, dass die Interviewerin ihr Verhalten der konkreten Gesprächssituation anpasste, anstatt durchgängig dieselben Fragen zu formulieren.

Für die ARJS liegt bereits eine Studie vor, die Unterschiede zwischen freien Berichten und Interviews überprüfte. Sporer und Walther (2006) fanden im Gegensatz zu den vorliegenden Befunden keinen Unterschied in der Validität der Skala Fehler und sozial Unerwünschtes zwischen freien Berichten und Interviews. Innerhalb der Interviews wurde zusätzlich die spezifische Fragetechnik variiert. Die

Hälfte der Interviews erfolgte anhand von fünf standardisierten offenen Fragen. Die andere Hälfte der Interviews wurde anhand von 20 Fragen durchgeführt, die direkt aus den ARJS abgeleitet wurden. Dabei wurde unter anderem gezielt nach den vier Merkmalen der Skala Fehler und sozial Unerwünschtes gefragt. Es ergab sich eine signifikante Wechselwirkung zwischen der Fragetechnik und dem Wahrheitsstatus für die Skala Fehler und sozial Unerwünschtes. Ein einfacher Haupteffekt des Wahrheitsstatus war nur bei der ARJS-Befragung festzustellen. Wahre Aussagen erzielten höhere Beurteilungen hinsichtlich der Skala Fehler und sozial Unerwünschtes als erfundene. Hingegen waren bei der standardisierten offenen Befragung die Mittelwerte wahrer und erfundener Aussagen vergleichbar. Dies ließ sich darauf zurückführen, dass wahr aussagende Personen von einer ARJS-Befragung profitierten. Sozial Unerwünschtes wurde bei wahrheitsgemäßen Schilderungen eher berichtet, wenn gezielt danach gefragt wurde als wenn allgemeine Fragen gestellt wurden. Erfundene Aussagen erzielten hingegen bei beiden Fragetechniken vergleichbare Beurteilungen. Dieser Befund unterstützt die Argumentation, dass lügende Personen es gezielt vermeiden, Fehler und sozial Unerwünschtes zu schildern. Zudem lässt sich aufgrund der Befunde von Sporer und Walther schlussfolgern, dass unterschiedliche Fragetechniken die Aussagequalität beinflussen können. Dies steht im Einklang mit Forschungsbefunden, die den moderierenden Einfluss von Fragetechniken auf die Validität der CBCA untersuchten (Craig et al., 1999; Davies, Westcott, & Horan, 2000; Hershkowitz, Lamb, Sternberg, & Esplin, 1997).

In der Untersuchung von Sporer und Walther (2006) waren sämtliche Interview-Fragen vollständig standardisiert. Hingegen wurden sie in der vorliegenden Studie der jeweiligen Aussage angepasst. Zudem verwendete die Interviewerin verschiedene Gesprächsführungstechniken, beispielsweise das Paraphrasieren vorangegangener Äußerungen der Probandin, die zu einer positiven Gesprächsatmosphäre beitragen können (vgl. Davies et al., 2000; Santilla et al., 2000, zum Einfluss des Interviewstils auf die CBCA). Daher ist

anzunehmen, dass interaktive Aspekte bei der vorliegenden Untersuchung stärker im Vordergrund standen als bei Sporer und Walther (2006). Dies könnte erklären, warum sich hier univariate Unterschiede zwischen wahren und erfundenen Interviews hinsichtlich der ARJS-Skala Fehler und sozial Unerwünschtes zeigten, während sie in der Untersuchung von Sporer und Walther (2006) nur feststellbar waren, wenn direkt danach gefragt wurde.

Effekte der Valenz der geschilderten Ereignisse

Schließlich wurde der Einfluss der Valenz der geschilderten Ereignisse auf die inhaltlichen Aussagemerkmale untersucht. Es wurde angenommen, dass die ARJS-Beurteilungen für negative Ereignisse höher ausfallen als für positive. Den Stimuluspersonen wurden verschiedene bedeutsame Lebensereignisse zur Auswahl vorgegeben. Dadurch resultierten sowohl Aussagen über vermeintlich positive (z.B. Hochzeit, Geburt eines Kindes) als auch über negative (z.B. Tod einer nahestehenden Person, Unfall) Themen. Dennoch ist zu berücksichtigen, dass ein und dasselbe Lebensereignis in Abhängigkeit von den spezifischen Umständen unterschiedlich erlebt werden kann. Beispielsweise kann das vermeintlich positive Erlebnis der Geburt eines Kindes wegen medizinischer Komplikationen durchaus als traumatisch empfunden werden. Daher gaben die Raterinnen für jede Aussage Beurteilungen hinsichtlich der Valenz des geschilderten Ereignisses ab, die als Grundlage der statistischen Analysen verwendet wurden. Nach dem Polyanna-Prinzip (vgl. Matlin, 2004) wäre zu erwarten gewesen, dass bei freier Themenwahl mehr positive als negative autobiographische Ereignisse berichtet werden. Diese Annahme wurde jedoch bereits durch andere Untersuchungen in Frage gestellt, die eine Gleichverteilung positiver und negativer Ereignisse berichteten (Sporer & Küpper, 1995, 2004; Sporer & Sharman; 2006). In diesen Studien wurden als Stimulusmaterial schriftliche Darstellungen verwendet und die Valenz der geschilderten Ereignisse über Selbstbeurteilungen erfasst. Hingegen erfolgten die Aussagen in der vorliegenden Untersuchung mündlich und deren Valenz wurde über

Fremdbeurteilungen erfasst. Dabei zeigte sich, dass im Gegensatz zu den Implikationen des Polyanna-Prinzips deutlich mehr negative als positive Ereignisse geschildert wurden. Mittels Mediandichotomisierung als positiv aufzufassende Aussagen erzielten auf der siebenstufigen Ratingskala einen Mittelwert von M = 4.13 (SD = 1.42; negative Ereignisse M = 2.11, SD = 0.45). Positive Ereignisse wurden also vor dem Hintergrund möglicher Beurteilungen eher als neutral bewertet. Trotz der vergleichsweise geringen Unterschiede in der Valenz ergaben sich jedoch erwartungsgemäße Effekte auf die inhaltliche Aussagequalität.

Für beide Aussageformen wiesen negative Ereignisse eine höhere Aussagequalität hinsichtlich der Skalen Memorieren und Gedächtnis, nonverbale und verbale Interaktionen sowie persönliche Signifikanz auf als positive. Für die freien Berichte ergaben sich zudem erwartungsgemäße Mittelwertsunterschiede für die Skalen Klarheit und Lebendigkeit, räumliche Details sowie Fehler und sozial Unerwünschtes. Die Ergebnisse stehen im Einklang mit den Befunden von Barnier et al. (2005), die ebenfalls höhere ARJS-Beurteilungen bei negativen im Vergleich zu positiven Ereignissen berichteten.

Nach Conway und Pleydell-Pearce (2000) ist davon auszugehen, dass negative Ereignisse mit Ängsten und dem Wunsch des Vergessens einhergehen. Infolgedessen sollte es eine höhere kognitive Anstrengung erfordern, sich an negative Ereignisse als an positive zu erinnern bzw. diese zu erfinden (de Vries, Blando & Walker, 1995; siehe auch Barnier et al., 2005). Wenn Personen mehr kognitive Anstrengung investieren, könnte dies in einer höheren Aussagequalität negativer im Vergleich zu positiven Ereignissen resultieren.

Zudem verwies Christianson (1992) darauf, dass die Zugänglichkeit negativer Erinnerungen von der Form der Befragung abhängig ist. Er argumentiert, dass im Rahmen von freien Berichten negative Ereignisse schlechter zugänglich sind als neutrale. Werden jedoch verschiedene Arten von Abrufhinweisen vorgegeben, wie es beispielsweise im Rahmen von Interviews geschieht, sollten

Unterschiede im Abruf emotionaler und neutraler Ereignisse verschwinden (Christianson, 1992). In der vorliegenden Untersuchung ergaben sich jedoch sowohl für die freien Berichte als auch für die Interviews Qualitätsunterschiede in Abhängigkeit von der Valenz der Ereignisse. Zudem zeigte sich, dass die negativen Ereignisse von höherer persönlicher Signifikanz waren und vermutlich daher häufiger memoriert wurden als positive Ereignisse. Dies würde erklären, warum die in der vorliegenden Untersuchung geschilderten negativen Ereignisse unabhängig von der Befragungsform gut zugänglich waren.

Insgesamt beeinflusste die Valenz des geschilderten Ereignisses die Fremdbeurteilungen der Aussagequalität. Dennoch moderierte die Valenz des geschilderten Ereignisses nicht den Zusammenhang zwischen dem objektiven Wahrheitsstatus und der Aussagequalität. Die ARJS bildeten unabhängig von der Valenz Unterschiede zwischen wahren und erfundenen Aussagen ab. Auch war kein korrelativer Zusammenhang zwischen der Eignung der Aussagen für die ARJS-Analyse und deren Valenz festzustellen. Daher lässt sich schlussfolgern, dass die Validität der inhaltlichen Glaubhaftigkeitsanalyse nicht auf negative Ereignisse begrenzt ist. Der Gültigkeitsanspruch der ARJS scheint im Gegensatz zu den Annahmen einiger SVA-Autoren sowohl für negative als auch für positive Ereignisse gegeben zu sein (vgl. Landry & Brigham, 1992; Steller, 1989).

Weiterführender Forschungsbedarf und praktische Implikationen

Aufgrund der vorliegenden Untersuchungsbefunde ist anzunehmen, dass die ARJS eine reliable und valide Beurteilung der Aussagequalität erlauben. Die Befunde zur Validität der ARJS sind ermutigend und durchaus mit dem Forschungsstand zur Validität der CBCA vergleichbar. Allerdings wurden wie bereits erwähnt in keiner Laborstudie alle CBCA-Merkmale untersucht, weil sie sich teilweise nicht auf das verwendete Stimulusmaterial anwenden ließen. Infolgedessen ist der Forschungsstand zur Validität der CBCA unvollständig und unübersichtlich. Der RÜ-Ansatz wiederum wurde zur Erklärung der individuellen

Wirklichkeitskontrolle von Gedächtnisinhalten entwickelt. Daher scheinen Modifikationen notwendig zu sein, um diesen Ansatz auf den Bereich der Entdeckung von Täuschung zu übertragen. Zudem steht eine einheitliche Operationalisierung der RÜ-Merkmale noch aus. Im Gegensatz dazu erlaubt die Formulierung der ARJS, die Aussageanalyse auf eine Vielzahl unterschiedlicher Themen anzuwenden. Darüber hinaus gewährleistet die standardisierte Durchführung eine direkte Vergleichbarkeit der Befunde einzelner Studien (vgl. Tabelle 3.13).

Des Weiteren eröffnete die Analyse von Effekten der Aussageform, Vorbereitungszeit und Valenz der geschilderten Ereignisse auf die Aussagequalität interessante Forschungs- und Anwendungsperspektiven, die im Folgenden herausgearbeitet werden. Auch wenn bei der Gesamtanalyse kein Effekt der Aussageform auf die Validität der ARJS vorlag, ergaben sich Unterschiede hinsichtlich der Skala Fehler und sozial Unerwünschtes. Erwartungsgemäße Unterschiede zwischen wahren und erfundenen Aussagen waren nur für die Interviews, nicht jedoch für die freien Berichte nachweisbar. Daher sollte der Einfluss von Interaktivität auf die Validität von Qualitätsmerkmalen, die aus dem Impression-Management-Ansatz abgeleitet wurden, in weiteren Studien gezielt überprüft werden. Aufgrund der vorliegenden Befunde ließe sich spekulieren, dass eine angenehme Gesprächsatmosphäre zur Validität der Skala Fehler und sozial Unerwünschtes beiträgt. Die gezielte Variation des Interviewerstils könnte weiteren Aufschluss über die differentielle Validität solcher Aussagemerkmale geben (vgl. Davies et al., 2000; Santilla et al., 2000; Vrij et al., 2006).

Wenn Personen gegenüber der Polizei oder bei Gericht aussagen, kann dies bedeutsame rechtliche Konsequenzen haben. Daher ist für praktische Anwendungen der Glaubhaftigkeitsanalyse zu vermuten, dass sich Personen auf ihre Aussage vorbereiten. Nach den vorliegenden Befunden wäre infolgedessen eine erhöhte Aussagequalität zu erwarten. Allerdings verschwand dieser Effekt,

wenn zwischen der Vorbereitung und der Aussage selbst eine Woche verging. Ließe sich dieses Ergebnis replizieren, könnte es aus Anwenderperspektive sinnvoll sein, nicht nur abzuschätzen ob, sondern auch wann eine Aussage vorbereitet wurde. In weiteren Studien wäre zudem auszudifferenzieren, wie viel Zeit zwischen der Vorbereitung einer Aussage und der Aussage selbst vergehen kann, damit ein Effekt auf die Aussagequalität erhalten bleibt. Ebenso ist zu beachten, dass mit zunehmendem Informationsgehalt von Aussagen externe Beweismittel an Bedeutung gewinnen können, weil sie beispielsweise Alibis bestätigen oder widerlegen. Daher wäre es auch sinnvoll zu untersuchen, welche konkreten Fragen Personen bei einer Vernehmung erwarten und vorbereiten.

Schließlich wurden bislang nur Aussagen von Erwachsenen hinsichtlich der ARJS analysiert. Dabei zeigten sich wiederholt Deckeneffekte für die Skala Realismus und logische Struktur. Infolgedessen war es nicht möglich Effekte des Wahrheitsstatus hinsichtlich dieses Merkmals aufzudecken. Möglicherweise lässt sich für die entsprechende Skala eher Variabilität beobachten, wenn Aussagen anderer Personengruppen analysiert werden. Daher erscheint es sinnvoll, die ARJS beispielsweise auch auf kindliche Aussagen anzuwenden, um die Gütekriterien für diese Skala erneut zu überprüfen (Sporer, 2004).

Glaubhaftigkeitsbegutachtungen werden vor allem dann angefordert, wenn widersprüchliche Aussagen zu einem inkriminierten Delikt vorliegen und externe Beweismittel fehlen. In der rechtspsychologischen Praxis betrifft dies insbesondere Fälle des sexuellen Missbrauchs, also von der Valenz her negative Ereignisse. Es ist jedoch nicht auszuschließen, dass auch andere Anwendungsgebiete für die inhaltliche Glaubhaftigkeitsanalyse erschlossen werden könnten. Im strafrechtlichen Bereich könnten beispielsweise Aussagen über Alibis überprüft werden. Auch bei Asylanträgen müssen die Entscheidungsträger oft aufgrund von Erlebnisberichten, die erlebnisbasiert oder erfunden sein können, Fälle entscheiden. Zudem wäre eine Anwendung der inhaltlichen Aussageanalyse im Familienrecht denkbar. Gegensätzliche Aussagen

sind auch bei Sorgerechtsstreitigkeiten oftmals nicht objektivierbar. So könnte die inhaltliche Glaubhaftigkeitsanalyse dazu beitragen aufzudecken, ob ein Elternteil tatsächlich Zeit mit seinen Kindern verbringt, oder nur vorgibt dies zu tun. Schließlich könnte es für Versicherungsunternehmen nützlich sein, Schadensansprüche auf ihre Glaubhaftigkeit hin überprüfen. Auf jeden Fall weisen die vorliegenden Befunde darauf hin, dass die inhaltliche Glaubhaftigkeitsanalyse auch bei verhältnismäßig positiven (oder zumindest neutralen) Ereignissen valide ist. Dieser Aspekt wurde in der Forschung zur inhaltlichen Glaubhaftigkeitsanalyse bislang zu sehr vernachlässigt. Zudem sollte die ARJS-Analyse nicht beschränkt auf die Vernehmung von Opfern und Zeugen anwendbar sein, sondern auch die Beschuldigtenvernehmung unterstützen können (vgl. Porter & Yuille, 1995). Aus historischer Perspektive sollte man letztlich nicht vergesssen, dass inhaltliche Merkmale noch lange vor den Psychologen von Juristen zur Analyse von Aussagen herangezogen wurden (vgl. Sporer, 1982)

Zusammenfassend haben sich die ARJS in der vorliegenden Untersuchung als reliabel und valide erwiesen. Zudem scheinen Vorbereitung und die Valenz des geschilderten Ereignisses die Aussagequalität zu beeinflussen, ohne die Validität der ARJS einzuschränken. Befunde zur Reliabilität und Validität von Glaubhaftigkeitsmerkmalen sind jedoch stets vor dem Hintergrund des verwendeten Stimulusmaterials zu interpretieren. Um die Nützlichkeit der ARJS-Aussageanalyse einzuschätzen, ist es notwendig zu überprüfen, wie offensichtlich die Unterschiede zwischen wahren und erfundenen Aussagen sind. Daher wurde ein weiteres Experiment durchgeführt, in dem dieselben Aussagen von Laien hinsichtlich ihrer Glaubhaftigkeit eingeschätzt wurden. Zusätzlich wurde überprüft, ob eine ökonomischere Kurzfassung der ARJS ebenfalls dazu geeignet ist, wahre von erfundenen Aussagen zu unterscheiden.

Literatur

Akehurst, L., Bull, R., Vrij, A., & Köhnken, G. (2004). The effects of training professional groups and lay persons to use criteria-based content analysis to detect deception. Applied Cognitive Psychology, 18, 877-891.

Akehurst, L., Köhnken, G., & Höfer, E. (2001). Content credibility of accounts derived from live and video presentations. Legal and Criminological Psychology, 6, 65-83.

Akehurst, L., Köhnken, G., Vrij, A., & Bull, R. (1996). Lay persons' and police officers' beliefs regarding deceptive behaviour. Applied Cognitive Psychology, 10, 461-471.

Alonso-Quecuty, M. L. (1992). Deception detection and reality monitoring: A new answer to an old question? In F. Lösel, D. Bender, & T. Bliesener (Eds.), Psychology and law: International perspectives (pp. 328-332). Berlin: de Gruyter.

Anderson, J. R. (1983). The architecture of cognition. Cambridge: Harvard University Press.

Anderson, S. J., & Conway, M. A. (1993). Investigating the structure of autobiographical memories. Journal of Experimental Psychology: Learning, Memory, and Cognition, 1, 1178-1196.

Anderson, S. J., & Conway, M. A. (1997). Representation of autobiographical memories. In M. A. Conway (Ed.), Cognitive models of memory (pp. 217-246). Hove: Psychology Press.

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

Barnier, A. J., Sharman, S. J., McKay, L., & Sporer, S. L. (2005). Discriminating adults' genuine, imagined, and deceptive accounts of positive and negative childhood events. Applied Cognitive Psychology, 19, 985-1001.

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., & Elaad, E. (2003). The validity of psychophysiological detection of information with the guilty knowledge test: A meta-analytic review. Journal of Applied Psychology, 88, 131-151.

BGH 1 StR 156/98 - Urteil vom 17.12.1998, Landesgericht Mannheim. Aufgerufen im Juni 2007, von: http://www.hrr-strafrecht.de/hrr/1/98/1-156-98.php3.

BGH 1 StR 618/98 - Urteil vom 30.07.1999, Landesgericht Ansbach. Aufgerufen im Juni 2007, von http://www.hrr-strafrecht.de/hrr/1/98/1-618-98.php3.

Bohanek, J. G., Fivush, R., & Walker, E. (2005). Memories of positive and negative emotional events. Applied Cognitive Psychology, 19, 51-66.

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. Personality and Social Psychology Review, 10, 214-234.

Bortz, J. (1999). Statistik für Sozialwissenschaftler (5. Auflage). Berlin: Springer.

Bower, G. H., & Hilgard, E. H. (1983). Theorien des Lernens. Stuttgart: Klett-Cotta.

Brewer, W. F. (1986). What is autobiographical memory? In D. C. Rubin (Ed.), Autobiographical memory (pp. 25-49). Cambridge: Cambridge University Press.

Brewer, W. F. (1988). Memory for randomly sampled autobiographical events. In U. Neisser & E. Winograd (Eds.), Remembering reconsidered: Ecological and traditional approaches to the study of memory (pp. 21-90). Cambridge: Cambridge University Press.

Brewer, W. F. (1996). What is recollective memory? In D. C. Rubin (Ed.), Remembering our past: Studies in autbiographical memory (pp. 19-66). New York: Cambridge University Press.

Brown, N. R., & Schopflocher, D. (1998). Event cueing, event clusters, and the temporal distribution of autobiographical memories. Applied Cognitive Psychology, 12, 305-319.

Buck, J. A., Warren, A. R., Betman, S., & Brigham, J. C. (2002). Age differences in criteria-based content analysis scores in typical child sexual abuse interviews. Applied Developmental Psychology, 23, 267-283.

Christianson, S.-A. (1992). Emotional stress and eyewitness memory: A critical review. Psychological Bulletin, 112, 284-309.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale: Lawrence Erlbaum.

Colwell, K., Hiscock, C. K., & Memon, A. (2002). Interviewing techniques and the assessment of statement credibility. Applied Cognitive Psychology, 16, 287-300.

Conway, M. A. (1990). Autobiographical memory: An introduction. Milton Keynes, UK: Open University Press.

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. Psychological Review, 107, 261-288.

Craig, R. A., Scheibe, R., Raskin, D. C., Kircher, J. C., & Dodd, D. H. (1999). Interviewer questions and content analysis of children's statements of sexual abuse. Applied Developmental Science, 3, 77-85.

D'Argembeau, A., Comblain, C. & van der Linden, M. (2003). Phenomenal characteristics of autobiographical memories for positive, negative and neutral events. Applied Cognitive Psychology, 17, 281-294.

Davies, G. M., Westcott, H. L., & Horan, N. (2000). The impact of questioning style on the content of investigative interviews with suspected child sexual abuse victims. Psychology, Crime, and Law, 6, 81-97.

DePaulo, B. M. & Friedman, H. S. (1998). Nonverbal communication. In D. Gilbert, S. T. Fiske, & G. Lidzey (Eds.), Handbook of social psychology (4th ed., Vol. 2, pp. 3-40). New York: Random House.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. Psychological Bulletin, 129, 74-112.

Destun, L. M., & Kuiper, N. A. (1999). Phenomenal characteristics associated with real and imagined events: The effects of event valence and absorption. Applied Cognitive Psychology, 13, 175-186.

DeVries, B., Blando, J. A., & Walker, L. J. (1995). An exploratory analysis of the content and structure of the life review. In B. K. Haight & J. D. Webster (Eds.), The art and science of reminiscing (pp. 123-137). Washington, DC: Taylor & Francis.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. Psychological Methods, 1, 170-177.

Elaad, E., & Ben-Shakhar, G. (1991). Effects of mental countermeasures on psychophysiological detection of information. International Journal of Psychophysiology, 11, 99-108.

Fiedler, K. (1989a). Lügendetektion aus alltagspsychologischer Sicht. Psychologische Rundschau, 40, 127-140.

Fiedler, K. (1989b). Suggestion and credibility: Lie detection based on content-related cues. In V. A. Gheorghin, P. Netter, H. J. Eysenck, & R. Rosenthal (Eds.), Suggestion and suggestiblity: Theory and research (pp. 323-335). Berlin: Springer.

Fisher, R. P., & Geiselman, R. E. (1992). Memory-enhancing techniques for investigative interviewing: The cognitive interview. Springfield, Illinois: Charles C. Thomas.

Fiske, S. T., & Taylor, S. E. (1991). Social cognition. New York: McGraw-Hill.

Geiselman, R. E., Fisher, R. P., MacKinnon, D. P., & Holland, H. L. (1986). Enhancement of eyewitness memory with the cognitive interview. American Journal of Psychology, 99, 385-401.

Grice, H. P.(1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), Syntax and semantics (pp. 41-58). New York: Academic Press.

Gödert, H. W., Gamer, M., Rill, H. G., & Vossel, G. (2005). Statement validity assessment: Inter-rater reliability in the mock-crime paradigm. Legal and Criminological Psychology, 10, 225-245.

Granhag, P. A., Strömwall, L. A., & Landström, S. (2006). Children recalling an event repeatedly: Effects on RM and CBCA scores. Legal and Criminological Psychology, 11, 81-98.

Hastie, R. (1981). Schematic principles in human memory. In E. T. Higgins, C. P. Hermann, & M. P. Zanna (Eds.), Social cognition: The ontario symposium (Vol. 1, pp. 39-88). Hillsdale, NJ: Erlbaum.

Hernandez-Fernaud, E., & Alonso-Quecuty, M. (1997). The cognitive interview and lie detection: A new magnifying glass for Sherlock Holmes? Applied Cognitive Psychology, 11, 55-68.

Hershkowitz, I., Lamb, M. E., Sternberg, K. J., & Esplin, P. W. (1997). The relationships among interviewer utterance type, CBCA scores and the richness of children´s responses. Legal and Criminological Psychology, 2, 169-176.

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., Raskin, D. C., & Kircher, J. C. (1994). Mental and physical countermeasures reduce the accuracy of the concealed knowledge test. Journal of Applied Psychology, 79, 252-259.

Horowitz, S. W., Lamb, M. E., Esplin, P. W., Boychuk, T. D., Krispin, O., & Reiter-Lavery, L. (1997). Reliability of criteria-based content analysis of child witness statements. Legal and Criminological Psychology, 2, 11-21.

Johnson, M. K. (1985). The origin of memories. In P. C. Kendall (Ed.), Advances in cognitive-behavioral research and therapy (Vol. 4, pp. 1-27). New York: Academic Press.

Johnson, M. K. (1988). Reality monitoring: An experimental phenomenological approach. Journal of Experimental Psychology: General, 117, 390-394.

Johnson, M. K. (2006). Memory and reality. American Psychologist, 61, 760-771.

Johnson, M. K., Bush, J. G., & Mitchell, K. J. (1998). Interpersonal reality monitoring: Judging the sources of other people's memories. Social Cognition, 16, 199-224.

Johnson, M. K., Foley, M. A., Suengas, A. G., & Raye, C. L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. Journal of Experimental Psychology: General, 117, 371-376.

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

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. Psychological Review, 88, 67-85.

Kennedy, Q., Mather, M., & Carstensen, L. L. (2004). The role of motivation in the age-related positivity effect in autobiographical memory. Psychological Science, 15, 208-214.

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

Köhnken, G., Schimossek, E., Aschermann, E., & Höfer, E. (1995). The cognitive interview and the assessment of the credibility of adults´ statements. Journal of Applied Psychology, 80, 671-684.

Köhnken, G., & Wegener, H. (1985). Zum Stellenwert des Experiments in der forensischen Aussagepsychologie. Zeitschrift für experimentelle und angewandte Psychologie, 32, 104-119.

Küpper, B., & Sporer, S. L. (1995). Beurteilerübereinstimmung bei Glaubwürdigkeitsmerkmalen: Eine empirische Studie. In G. Bierbrauer (Ed.),

Verfahrensgerechtigkeit: Rechtspsychologische Forschungsbeiträge für die Justizpraxis (S. 187-213). Köln: Verlag Dr. Otto Schmidt KG.

Lakhani, M., & Taylor, R. (2003). Beliefs about the cues to deception in high- and low-stake situations. Psychology, Crime, and Law, 9, 357-368.

Lamb, M. E., Sternberg, K. J., Esplin, P. W., Hershkowitz, I., Orbach, Y., & Hovav, M. (1997). Criterion-based content analysis: A field validation study. Child Abuse & Neglect, 21, 255-264.

Landry, K., & Brigham, J. C. (1992). The effect of training in criteria-based content analysis on the ability to detect deception in adults. Law and Human Behavior, 16, 663-675.

Larsen, S. F. (1998). What is it like to remember? On phenomenal qualities of memory. In C. P. Thomson, D. J. Herrmann, D. Bruce, J. D. Read, D. G. Payne, & M. P. Toglia (Eds.), Autobiographical memory: Theoretical and applied perspectives (pp. 163-190). Mahwah, New Jersey: Lawrence Erlbaum.

Larsson, A. S., & Granhag, P. A. (2005). Interviewing children with the cognitive interview: Assessing the reliability of statements based on observed and imagined events. Scandinavian Journal of Psychology, 46, 49-57.

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. Thousand Oaks, CA: Sage.

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. Psychology, Crime, and Law, 11, 99-122.

Matlin, M. W. (2004). Pollyanna principle. In R. Pohl (Ed.), Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory (pp. 255-271). London: Psychology Press.

MAXQDA 2 [Computer software]. (2004). Marburg, Germany: Verbi GmBH.

McGinnis, D., & Roberts, P. (1996). Qualitative characteristics of vivid memories attributed to real and imagined experiences. American Journal of Psychology, 109, 59-77.

Miller, G. R., & Stiff, J. B. (1993). Deceptive Communication. Newbury Park, Californa: Sage Publications.

Mullen, B. (1989). Advanced basic meta-analysis. Hillsdale, New Jersey: Lawrence Erlbaum.

Niehaus, S. (2001). Zur Anwendbarkeit inhaltlicher Glaubhaftigkeitsmerkmale bei Zeugenaussagen unterschiedlichen Wahrheitsgehaltes: Eine Simulationsstudie mit kindlichen Verkehrsunfallopfern. Frankfurt am Main: Peter Lang.

Niehaus, S. (2003). Diskriminationsfähigkei der merkmalsorientierten Inhaltsanalyse bei teilweise erlebnisbasierten Falschaussagen. Praxis der Rechtspsychologie, 13, 309-328.

Parker, A. D., & Brown, J. (2000). Detection of deception: Statement validity analysis as a means of determining truthfulness or falsity of rape allegations. Legal and Criminological Psychology, 5, 237-259.

Pohl, R. (2007). Das autobiographische Gedächtnis: Die Psychologie unserer Lebensgeschichte. Stuttgart: Kohlhammer.

Pontari, B. A., & Schlenker, B. R. (2000). The influence of cognitive load on self-presentation: Can cognitive busyness help as well as harm social performance? Journal of Personality and Social Psychology, 78, 1092-1108.

Porter, S., & Yuille, J. C. (1996). The language of deceit: An investigation of the verbal clues to deception in the interrogation context. Law and Human Behavior, 20, 443-459.

Porter, S., Yuille, J. C., & Lehman, D. R. (1999). The nature of real, implanted and fabricated memories for emotional childhood events: Implications for the recovered memory debate. Law and Human Behavior, 23, 517-537.

Raskin, D. C., & Esplin, P. W. (1991). Assessment of children´s statements of sexual abuse. In J. Doris (Ed.), The suggestibility of children´s recollections (pp. 153-164). Washington, DC: American Psychological Association.

Rassin, E., & van der Sleen, J. (2005). Characteristics of true versus false allegations of sexual offences. Psychological Reports, 97, 589-598.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d and r. Law and Human Behavior, 29, 615-620.

Rojahn, K., & Pettigrew, T. F. (1992). Memory for schema-relevant information: A meta-analytic resolution. British Journal of Psychology, 31, 81-109.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), The handbook of research synthesis (pp. 231-244). New York: Russell Sage Foundation.

Rosenthal, R. (1995). Methodology. In A. Tesser (Ed.), Advanced social psychology (pp. 17-49). Boston: McGraw-Hill.

Santtila, P., Roppola, H., Runtti, M., & Niemi, P. (2000). Assessment of child witness statements using criteria-based content analysis (CBCA): The effects of age, verbal ability, and interviewer´s emotional style. Psychology, Crime, and Law, 6, 159-179.

Schaefer, A., & Philippot, P. (2005). Selective effects of emotion on the phenomenal characteristics of autobiographical memories. Memory, 13, 148-160.

Schlenker, B. R., & Weigold, M. F. (1992). Interpersonal processes involving impression regulation and management. Annual Review of Psychology, 43, 133-168.

Sporer, S. L. (1982). A brief history of the psychology of testimony. Current Psychological Reviews, 2, 323-340.

Sporer, S. L. (1996/1998/2004). The Aberdeen Report Judgment Scales (ARJS). Definitions and answer sheets. Unpublished Questionnaire. University of Aberdeen, Scotland; University of Giessen, Germany.

Sporer, S. L. (1997a). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experiences events. Applied Cognitive Psychology, 11, 373-397.

Sporer, S. L. (1997b). Realitätsüberwachungskriterien und forensische Glaubwürdigkeitskriterien im Vergleich. In L. Greuel, T. Fabian, & M. Stadler (Eds.), Psychologie der Zeugenaussage (S. 71-85). München: Psychologie Verlags Union.

Sporer, S. L. (1998, March). Detecting deception with the Aberdeen Report Judgment Scales (ARJS): Theoretical development reliability and validity. Paper presented at the Biennial Meeting of the American Psychology-Law Society, Redondo Beach, CA.

Sporer, S. L. (2004). Reality monitoring and detection of deception. In P. A. Granhag & L. Strömwall (Eds.), Deception detection in forensic contexts (pp. 64-102). Cambridge: University Press.

Sporer, S. L., & Burghardt, S. E. (2004, March). Truth detection with the Aberdeen Report Judgment Scales: The role of planning and rehearsal. Paper presented at the Biennial Meeting of the American Psychology-Law Society, Phoenix, AZ.

Sporer, S. L., & Bursch, S. E. (1996, April). Detection of deception by verbal means: Before and after training. Paper presented at the 38. Tagung experimentell arbeitender Psychologen, Eichstaett, Germany.

Sporer, S. L., Bursch, S. E., Schreiber, N., Weiss, P. E., Höfer, E., Sievers, K., & Köhnken, G. (2000). Detecting deception with the Aberdeen Report Judgment Scales (ARJS): Inter-Rater Reliability. In A. Czerederecka, T. Jaskiewicz-Obydzinska, & J. Wojcikiewicz (Eds.), Forensic psychology and law (pp. 197-204). Krakow: Institute of Forensic Research Publishers.

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

Sporer, S. L., & Küpper, B. (2004). Fantasie und Wirklichkeit: Erinnerungsqualitäten von erlebten und erfundenen Geschichten. Zeitschrift für Psychologie, 212, 135-151.

Sporer, S. L., Samweber, M. C., & Stucke, T. S. (2000, March). Twisting the outcome: Discriminating distorted truths from factually experienced events. Paper presented at the Biennial Meeting of the American Psychology-Law Society, New Orleans; LA.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. Applied Cognitive Psychology, 20, 421-446.

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. Psychology, Public Policy, and Law, 13, 1-34.

Sporer, S. L., & Sharman, S. J. (2006). Should I believe this? Reality monitoring of accounts of self-experienced and invented recent and distant autobiographical events. Applied Cognitive Psychology, 20, 1-18.

Sporer, S. L., & Walther, A. (2006, March). Truth detection by content cues: General vs. specific questions. Paper presented at the Meeting of the American Psychology-Law Society in Petersburg, FL.

Sporer, S. L., & Zander, J. (2001, June). Nonverbal cues to deception: Do motivation and preparation make a difference? Paper presented at the 11th European Conference on Psychology and Law in Lisbon, Portugal.

Srull T. K., & Wyer, R. S. (1989). Person memory and judgement. Psychological Review, 96, 58-83.

Stangor, C., & McMillan, D. (1992). Memory of expectancy-congruent and expectancy-incongruent information: A review of the social and social developmental literatures. Psychological Bulletin, 111, 42-61.

Steller, M. (1988). Die vierte Phase der Aussagepsychologie. Forensia, 9, 23-28.

Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille (Ed.), Credibility Assessment (pp. 135-154). Deventer: Kluwer Academic Publishers.

Steller, M., & Köhnken, G. (1989). Criteria-based statement analysis. In D. C. Raskin (Ed). Psychological methods in criminal investigation and evidence (pp. 217-245). New York: Springer.

Steller, M., & Boychuk, T. (1992) Children as witnesses in sexual abusse cases: Investigative interview and assessment techniques. In H. Dent & R. Flin (Eds.), Children as witnesses (pp.47-71). Chichester: Wiley.

Steller, M., & Wellershaus, P. (1996). Information enhancement and credibility assessment of child statements: The impact of the cognitive interview technique on criteria-based content analysis. In G. Davies, S. Lloyd-Bostock, M. McMurran, & C. Wilson (Eds.), Psychology, law, and criminal justice: International developments in research and practice (pp. 118-127). Berlin: de Gruyter.

Steller, M., Wellershaus, P., & Wolf, T. (1992). Realkennzeichen in Kinderaussagen: Empirische Grundlage der Kriterienorientierten Aussageanalyse. Zeitschrift für Experimentelle und Angewandte Psychologie, 39, 151-170.

Stelzl, I. (2005). Fehler und Fallen der Statistik: Für Psychologen, Pädagogen und Sozialwissenschaftler. Münster: Waxmann.

Strömwall, L. A., Bengtsson, L., Leander, L., & Granhag, P. A. (2004). Assessing children's statements: The impact of a repeated experience on CBCA and RM ratings. Applied Cognitive Psychology, 18, 653-668.

Suengas, A. G., & Johnson, M. K. (1988). Qualitative effects of rehearsal on memories for perceived and imagined complex events. Journal of Experimental Psychology: General, 117, 377-389.

Taylor, R., & Vrij, A. (2000). The effects of varying stake and cognitive complexity on beliefs about the cues to deception. International Journal of Police Science and Management, 3, 111-123.

Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Aussagen. In U. Undeutsch (Ed.), Handbuch der Psychologie, Bd. 11., Forensische Psychologie (S. 26-181). Göttingen, Germany: Hogrefe.

Vrij, A. (2005). Criteria-based content analysis: A qualitative review of the first 37 studies. Psychology, Public Policy, and Law, 11, 3-41.

Vrij, A., Akehurst, L., & Knight, S. (2006). Police officers´, social workers' and the general public´s beliefs about deception in children, adolescents and adults. Legal and Criminological Psychology, 11, 297-312.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004a). Let me inform you how to tell a convincing story: CBCA and reality monitoring scores as a function of age, coaching, and deception. Canadian Journal of Behavioral Science-Revue, 36, 113-126.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004b). Detecting deceit via analyses of verbal and nonverbal behavior in children and adults. Human Communication Research, 30, 8-41.

Vrij, A., Edward, K., & Bull, R. (2001a). People's insight into their own behaviour and speech content while lying. British Journal of Psychology, 92, 373-389.

Vrij, A., Edward, K., & Bull, R. (2001b). Stereotypical verbal and nonverbal responses while deceiving others. Personality and Social Psychology Bulletin, 27, 899-909.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. Journal of Nonverbal Behavior, 24, 239-263.

Vrij, A., Evans, H., Akehurst, L., & Mann, S. (2004). Rapid Judgments in assessing verbal and nonverbal cues: Their potential for deception researchers and lie detection. Applied Cognitive Psychology, 18, 283-296.

Vrij, A., Kneller, W., & Mann, S. (2000). The effect of informing liars about Criteria-Based Content Analysis on their ability to deceive CBCA-raters. Legal and Criminological Psychology, 5, 57-70.

Wegener, H. (1997) Die Entwicklung der experimentellen Aussagepsychologie. In L. Greuel, T. Fabian & M. Stadler (Eds.), Psychologie der Zeugenaussage (S. 13-22). München: Psychologie Verlags Union.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. Personality and Social Psychology Bulletin, 25, 1115-1125.

Wirtz, M., & Caspar, F. (2002). Beurteilerübereinstimmung und Beurteilerreliabilität. Göttingen: Hogrefe.

Zaparniuk, J., Yuille, J. C., & Taylor, S. (1995). Assessing the credibility of true and false statements. International Journal of Law and Psychiatry, 18, 343-352.

Zuckerman, M., & Driver, R. E. (1985). Telling lies: Verbal and nonverbal correlates of deception. In A. W. Siegman & S. Feldstein (Eds.), Multichannel integrations of nonverbal behaviour (pp. 129-148). Hillsdale, NJ: Erlbaum.

Anhang 3a

Haupteffekt des Wahrheitsstatus auf die 13 ARJS-Skalen bei Kontrolle der Anzahl der Wörter

	Freie Berichte					Interviews				
Skalen	$M_{erfunden}$	M_{wahr}	$F(1,171)$	p	r	$M_{erfunden}$	M_{wahr}	$F(1,171)$	p	r
Realismus und logische Struktur	6.97	6.98	0.46	.500	.06	6.96	6.96	0.02	.879	.01
Klarheit und Lebendigkeit	5.12	5.42	5.82	**.017**	**.19**	5.44	5.77	11.04	**.001**	**.24**
Details	3.76	4.10	6.22	**.014**	**.18**	4.02	4.49	20.62	**.000**	**.33**
Räumliche Details	2.63	2.56	0.17	.685	-.03	3.33	3.45	0.46	.501	.05
Zeitliche Details	3.23	3.48	1.15	.285	.07	4.27	4.61	2.56	.111	.12
Sinneseindrücke	1.28	1.21	1.00	.319	-.07	1.33	1.29	0.32	.573	-.04
Emotionen und Gefühle	3.70	4.01	2.97	.087	.13	3.56	3.91	4.89	**.028**	**.17**
Gedanken	4.04	4.11	0.10	.753	.03	3.88	4.04	0.52	.470	.06
Memorieren und Gedächtnis	2.45	2.50	0.20	.657	.03	2.46	2.60	1.90	.170	.11
Nonverbale und verbale Interaktionen	2.93	2.88	0.11	.741	-.03	2.86	3.06	2.08	.151	.11
Komplikationen und ungewöhnliche Details	1.75	1.95	1.83	.178	.10	1.78	2.12	6.70	**.010**	**.19**
Fehler und sozial Unerwünschtes	1.97	1.95	0.02	.894	-.01	2.37	2.68	5.10	**.025**	**.17**
Persönliche Signifikanz	3.82	3.42	6.66	**.011**	**-.18**	3.72	3.36	5.48	**.020**	**-.17**

Anm. N = 176. Positive Partialkorrelationen r signieren eine höhere Ausprägung in wahren Berichten.

STUDIE 4

Inter-Rater-Reliabilität und Validität der Aberdeen Report Judgment Scales--Short Training Version--German

Im Rahmen der vorangegangenen Studie wurden Befunde zu den Hauptgütekriterien der Inter-Rater-Reliabilität und Validität für die Aberdeen Report Judgment Scales (ARJS, Sporer, 1996) vorgestellt. Für praktische Anwendungen der Glaubhaftigkeitsdiagnostik kommt jedoch auch dem Nebengütekriterium der Ökonomie wesentliche Bedeutung zu. Die in der vorangegangenen Studie durchgeführte ARJS-Aussageanalyse erforderte einen hohen Aufwand in Vorbereitung und Durchführung. Die Beurteilerinnen verfügten über umfangreiches Hintergrundwissen, absolvierten eine 3-monatige Schulung, und die Bearbeitung der einzelnen Transkripte dauerte trotz intensiver Übungseffekte durchschnittlich ungefähr eine Stunde. Für praktisch tätige Sachverständige erscheint ein solcher Aufwand zwar zumutbar, doch schließt er andere potenzielle Anwendergruppen aus. So könnte die Glaubhaftigkeitsanalyse beispielsweise auch für die Ermittlungstätigkeiten von Polizeibeamten (z.B. bei der Überprüfung von Alibis) oder von Sachbearbeitern in Versicherungsunternehmen bei vermutlichen Betrugsfällen nützlich sein. Doch auch für Richter und Staatsanwälte könnten Hilfestellungen zur Analyse der Glaubhaftigkeit von Aussagen von Interesse sein. Daher stellt sich die Frage, ob auch Beurteilerinnen ohne Vorkenntnisse im Bereich der Glaubhaftigkeitsdiagnostik von der inhaltlichen Aussageanalyse profitieren könnten. Um dies zu klären wurde in der vorliegenden Studie dasselbe Stimulusmaterial anhand einer Kurzform der ARJS, der sogenannten Aberdeen Report Judgment Scales--Short Training Version--German (ARJS-STV-G) beurteilt.

Ziele und Hypothesen

Für die vorliegende Untersuchung wird angenommen, dass eine Kurzform der ARJS, die ARJS-STV-G, ebenfalls Qualitätsunterschiede zwischen wahren und

erfundenen Aussagen abbildet. Dabei wird erneut eine höhere inhaltliche Aussagequalität für wahre im Vergleich zu erfundenen Aussagen erwartet. Zudem soll untersucht werden, ob sich die in der vorangegangenen Studie beobachteten Effekte der Aussageform, der Vorbereitungszeit und der Valenz des geschilderten Ereignisses auch auf die Beurteilungen anhand der ARJS-STV-G auswirken.

Methode

Design der Untersuchung

Im Rahmen der vorliegenden Studie wurde das bereits zuvor verwendete Stimulusmaterial von acht unabhängigen Beurteilerinnen bewertet. Diese waren Studierende der Psychologie und verfügten über keinerlei Vorkenntnisse im Bereich der Glaubhaftigkeitsdiagnostik. Die Beurteilung der Aussagen erfolgte in zwei Phasen. Während der ersten Phase wurde die Glaubhaftigkeit der Aussagen naiv eingeschätzt, während der zweiten Phase anhand der ARJS-STV-G. Das Versuchsdesign ist Tabelle 4.1 zu entnehmen. Erneut beurteilten alle Raterinnen die Hälfte der freien Berichte sowie alle Interviews. Durch die Within-Subjects-Variation der Beurteilungsgrundlage resultierten für jeden freien Bericht zwei unabhängige naive und zwei unabhängige ARJS-STV-G-Beurteilungen. Für jedes Interview lagen hingegen jeweils vier unabhängige naive und vier ARJS-STV-G-Beurteilungen vor. Die Bearbeitungsreihenfolge wurde für Beurteilerinnen, die dieselben Aussagen unter derselben Bedingung bewerteten, ausbalanciert.

Phase 1: Naive Beurteilung

Während der ersten Phase gaben die Raterinnen naive Urteile hinsichtlich der Glaubhaftigkeit der einzelnen Aussagen ab. Für das subjektive Glaubhaftigkeitsurteil wurde die gleiche 10-stufige Skala wie in der vorangegangenen Studie verwendet. Erneut indizierten Werte von 1 bis 5, dass eine Aussage als erfunden, Werte von 6 bis 10 dass eine Aussage als wahr aufgefasst wurde. Zudem wurden die Raterinnen instruiert drei frei zu formulierende Urteilsbegründungen abzugeben und deren relative Wichtigkeit für

Tabelle 4.1

Versuchsdesign zur Aufteilung der N = 176 freien Berichte und N = 176 Interviews auf die acht Raterinnen

	Raterin								
Aussagen (je n = 88 pro Gruppe)	A	B	C	D	E	F	G	H	Summe
Freie Berichte									
B01.01-B03.12	Naiv	Naiv	---[a]	---[a]	ARJS-STV-G	ARJS-STV-G	---[a]	---[a]	2
B03.13-B06.08	ARJS-STV-G	ARJS-STV-G	---[a]	---[a]	Naiv	Naiv	---[a]	---[a]	2
B06.09-B09.04	---[a]	---[a]	Naiv	Naiv	---[a]	---[a]	ARJS-STV-G	ARJS-STV-G	2
B09.05-B11.16	---[a]	---[a]	ARJS-STV-G	ARJS-STV-G	---[a]	---[a]	Naiv	Naiv	2
Interviews									
C01.01-C03.12	Naiv	Naiv	Naiv	Naiv	ARJS-STV-G	ARJS-STV-G	ARJS-STV-G	ARJS-STV-G	4
C03.13-C06.08	ARJS-STV-G	ARJS-STV-G	ARJS-STV-G	ARJS-STV-G	Naiv	Naiv	Naiv	Naiv	4
C06.09-C09.04	Naiv	Naiv	Naiv	Naiv	ARJS-STV-G	ARJS-STV-G	ARJS-STV-G	ARJS-STV-G	4
C09.05-C11.16	ARJS-STV-G	ARJS-STV-G	ARJS-STV-G	ARJS-STV-G	Naiv	Naiv	Naiv	Naiv	4

Anm. N = 176 freie Berichte (B01.01-B11.16) und N = 176 Interviews (C01.01-C11.16). Naiv = naive Beurteilung während der ersten Untersuchungsphase; ARJS-STV-G = Beurteilung anhand der 17 ARJS-STV-G Merkmale während der zweiten Untersuchungsphase; ---[a] = keine Beurteilung; Summe = Anzahl unabhängiger Beurteilungen für die entsprechenden Aussagen.

das Glaubhaftigkeitsurteil zu vermerken. Schließlich wurde die subjektive Sicherheit hinsichtlich des Glaubhaftigkeitsurteils wie in der vorangegangenen Studie auf einer 5-stufigen Skala erhoben. Niedrige Werte verwiesen auf ein unsicheres, hohe auf ein sehr sicheres Urteil. Die naiven Beurteilungen sind jedoch nicht Gegenstand der vorliegenden Untersuchung, sondern wurden in Studie 5 analysiert.

Phase 2: Beurteilung anhand der ARJS-STV-G

Im Anschluss an die naive Beurteilungsphase wurden die Raterinnen in einer 2,5-stündigen Schulung von der Verfasserin über die Anwendung der ARJS-STV-G informiert. Dabei wurden schriftliche Definitionen der zu beurteilenden Glaubhaftigkeitsmerkmale vorgestellt und anhand von Beispielsätzen illustriert. Zudem wurde eine Übungsaussage beurteilt, um eine korrekte Anwendung der ARJS-STV-G zu gewährleisten.

Die Kurzform der ARJS wurde in englischer und spanischer Sprache verfasst (Cramer, 2005; Sporer & Masip, 2007) und für die vorliegende Untersuchung ins Deutsche übersetzt. Die ARJS-STV-G umfasst 17 Glaubhaftigkeitsmerkmale, die der Langform entnommen wurden. In der dritten Studie (vgl. Tabelle 3.1) wurde angegeben, welche Merkmale der Langform sich in der Kurzform wiederfinden. Von diesen Merkmalen entsprechen elf den Skalenbenennungen der Langform, die übrigen sechs entsprechen einzelnen Items der Langform. Neben überflüssigen Details (vgl. ARJS-Skala Details) werden spontane Korrekturen, Erinnerungslücken zugeben, negative Äußerungen über sich selbst (vgl. ARJS-Skala Fehler und sozial Unerwünschtes), sowie Komplikationen und ungewöhnliche Details als Glaubhaftigkeitsmerkmale vorgestellt. Demnach werden Merkmale, die aus dem Impression-Management-Ansatz abgeleitet wurden, auch bei der ARJS-STV-G relativ differenziert erfasst.

Sämtliche Merkmale sind auf 7-stufigen Skalen zu beurteilen, die durch passende Adjektive bipolar verankert sind (z.B. keine--viele, vage--präzise). Analog

der ARJS reflektieren geringe Werte, dass ein Glaubhaftigkeitsmerkmal nicht oder nur in geringer Qualität vorliegt. Hingegen verweisen hohe Werte auf eine hohe Quantität oder Qualität. Die Beurteilungen der 17 Glaubhaftigkeitsmerkmale werden anschließend zu drei Merkmalsgruppen integriert, die jeweils 3-stufig zu bewerten sind. Die einzelnen Stufen reflektieren eine geringe (0), mittlere (1) und starke Ausprägung (2) der einer Gruppe zugehörigen Merkmale. Für das abschließende Glaubhaftigkeitsurteil sind die drei Gruppenurteile unterschiedlich zu gewichten. Die Gruppe mit dem geringsten Gewicht umfasst die beiden globalen Merkmale persönliche Bedeutsamkeit und Klarheit und Lebendigkeit sowie die persönliche Signifikanz des Ereignisses. Merkmale, die sich auf die Detailliertheit der Aussage, Gedanken, Gedächtnisprozesse, sensorische Eindrücke, Gefühle und Interaktionen beziehen erhalten ein mittleres Gewicht. Am stärksten gewichtet werden selbstdarstellungsbasierte Merkmale sowie Komplikationen, ungewöhnliche und überflüssige Details. Für das abschließende Glaubhaftigkeitsurteil und die subjektive Sicherheit wurden dieselben 10- bzw. 5-stufigen Antwortskalen wie für die naiven Beurteilungen verwendet.

Ergebnisse

Einleitend werden deskriptive Kennwerte der ARJS-STV-G-Beurteilungen für das gesamte Stimulusmaterial dargestellt. Danach wird darauf eingegangen, wie reliabel die Glaubhaftigkeitsmerkmale beurteilt wurden. Schließlich werden Befunde zur Validität der ARJS-STV-G-Merkmale vorgestellt. Dabei werden Einflüsse des tatsächlichen Wahrheitsstatus, der Gelegenheit zur Vorbereitung, der Aussageform und der Valenz der geschilderten Ereignisse auf die Beurteilungen der Aussagequalität überprüft. Insgesamt wurden dabei die gleichen Auswertungsverfahren verwendet und statistischen Indikatoren abgeleitet wie im vorangegangenen Experiment. Ausführlichere Informationen zum Stimulusmaterial und zu den angewandten Formeln sind daher der

vorangegangenen Untersuchung zur Evaluation der ARJS zu entnehmen (Studie 3).

Deskriptive Analysen und interne Konsistenz der ARJS-STV-G

Um die Aussagequalität des verwendeten Stimulusmaterials hinsichtlich der 17 ARJS-STV-G-Merkmale zu beschreiben, wurden zunächst deren Mittelwerte, Standardabweichungen, Minima und Maxima ermittelt. Die in Tabelle 4.2 dargestellten Werte basieren auf den mittleren Beurteilungen durch alle acht Raterinnen. Die beiden Merkmale Realismus und logische Struktur sowie Klarheit und Lebendigkeit wiesen die höchsten Mittelwerte auf. Ebenso wurde die persönliche Signifikanz der Ereignisse recht hoch eingeschätzt. Im Gegensatz dazu wurden nur selten negative Äußerungen über sich selbst, spontane Korrekturen und sensorische Eindrücke im Stimulusmaterial vorgefunden. Die Merkmale bildeten jedoch trotz reduzierter Mittelwerte Variabilität zwischen den Aussagen ab.

Inter-Rater-Reliabilität der ARJS-STV-G

Zur Bestimmung der Inter-Rater-Reliabilitäten wurden paarweise Korrelationen zwischen den Raterinnen, die dieselben Aussagen beurteilten, berechnet. Diese Korrelationen wurden anschließend einer Fisher's Z-Transformation unterzogen, dann gemittelt und in r zurücktransformiert. Nachfolgende Analysen zur Validität der ARJS-STV-G basieren auf den über verschiedene Raterinnen gemittelten Beurteilungen. Um die Reliabilität dieser Gruppenurteile abzuschätzen, wurden die mittleren Korrelationen anhand der Spearman-Brown-Formel korrigiert. Im Folgenden werden die Befunde zur Inter-Rater-Reliabilität für beide Aussageformen getrennt dargestellt.

Tabelle 4.2
Mittelwerte und Standardabweichungen der ARJS-STV-G-Beurteilungen

	Freie Berichte				Interviews			
Merkmale	M	SD	Min	Max	M	SD	Min	Max
Negative Äußerungen über sich selbst	1.82	1.15	1.00	7.00	1.83	1.04	1.00	6.25
Spontane Korrekturen	1.48	1.04	1.00	7.00	1.62	0.91	1.00	5.75
Erinnerungslücken zugeben	1.88	1.40	1.00	7.00	3.12	1.74	1.00	7.00
Komplikationen	2.18	1.36	1.00	6.00	2.16	1.08	1.00	5.25
Ungewöhnliche Details	2.10	1.28	1.00	6.50	2.08	0.99	1.00	5.75
Überflüssige Details	2.87	1.54	1.00	7.00	3.19	1.09	1.00	6.25
Nonverbale und verbale Interaktionen	4.04	1.88	1.00	7.00	4.01	1.67	1.00	7.00
Emotionen und Gefühle	4.84	1.30	1.00	7.00	4.83	1.24	1.50	7.00
Sensorische Eindrücke	1.88	1.43	1.00	6.50	2.03	1.27	1.00	6.50
Gedächtnisprozesse und Memorieren	3.58	1.61	1.00	7.00	3.75	1.06	1.50	7.00
Gedanken	4.51	1.76	1.00	7.00	4.61	1.25	1.25	7.00
Räumliche Details	3.26	1.24	1.00	7.00	4.17	1.16	1.50	7.00
Zeitliche Details	4.06	1.52	1.00	7.00	5.20	1.09	1.75	7.00
Details	4.05	1.15	1.50	7.00	4.42	0.87	2.50	6.50
Persönliche Bedeutsamkeit	5.11	1.32	2.00	7.00	5.09	1.06	2.75	7.00
Klarheit und Lebendigkeit	5.42	1.07	2.00	7.00	5.63	0.81	3.00	7.00
Realismus und logische Struktur	5.73	0.96	3.00	7.00	5.91	0.67	3.50	7.00

Anm. N = 176 Freie Berichte und N = 176 Interviews. Werte basieren für die freien Berichte auf der mittleren Beurteilung durch jeweils zwei, für die Interviews durch jeweils vier unterschiedliche Raterinnen.

7-stufige Ratingskalen mit 1 = niedrige bis 7 = hohe Ausprägung.

Freie Berichte

Jeweils zwei der acht Raterinnen beurteilten dieselben freien Berichte anhand der ARJS-STV-G. Die Korrelationen zwischen diesen Raterinnenpaaren sind im Anhang 4a dokumentiert. Die mittleren paarweisen Korrelationen sind Tabelle 4.3 zu entnehmen. Diese variierten zwischen $\underline{r} = .71$ und $\underline{r} = .10$ mit der höchsten Reliabilität für das Merkmal nonverbale und verbale Interaktionen und der geringsten für spontane Korrekturen. Anhand einer Spearman-Brown-Korrektur für zwei Rater ließ sich zeigen, dass die Reliabilität der Gruppenurteile deutlich besser ausfiel. Für sechs Merkmale wurden zufrieden stellende Inter-Rater-Reliabilitäten von $\underline{r}_{SP(2)} \geq .60$ erzielt. Für vier Merkmale ergaben sich Spearman-Brown-korrigierte Werte von $.60 > \underline{r}_{SP(2)} \geq .50$ und für drei weitere von $.50 > \underline{r}_{SP(2)} \geq .40$. Die Merkmale spontane Korrekturen, ungewöhnliche Details, Gedächtnisprozesse und Memorieren sowie Realismus und logische Struktur wurden trotz der Zusammenfassung von zwei unabhängigen Beurteilungen nicht reliabel beurteilt. Für spontane Korrekturen und ungewöhnliche Details könnte dies an Bodeneffekten, für Realismus und logische Struktur an einem Deckeneffekt liegen.

Interviews

Alle acht Raterinnen beurteilten jeweils die Hälfte der Interviews anhand der ARJS-STV-G, wobei jeweils vier Raterinnen dieselben Aussagen beurteilten. Die paarweisen Korrelationen zwischen diesen Raterinnen befinden sich im Anhang 4b. Ebenso sind dort die mittleren Korrelationen der Raterinnenpaare, die dieselben Interviews beurteilten, aufgeführt. Diese wurden wiederum gemittelt, um die Inter-Rater-Reliabililität für alle Interviews abzuschätzen. Die Ergebnisse sind in Tabelle 4.3 den Inter-Rater-Reliabilitäten für die freien Berichte gegenübergestellt.

Tabelle 4.3

Inter-Rater-Reliabilitäten (Pearson-Korrelationen) der 17 ARJS-STV-G-Merkmale für die freien Berichte und die Interviews

	Freie Berichte		Interviews	
Merkmale	M(r)	$r_{SP(2)}$	M(r)	$r_{SP(4)}$
Negative Äußerungen über sich selbst	.41	.58	.31	.65
Spontane Korrekturen	.10	.18	.22	.53
Erinnerungslücken zugeben	.51	.67	.52	.81
Komplikationen	.26	.41	.24	.56
Ungewöhnliche Details	.18	.31	.22	.54
Überflüssige Details	.27	.43	.33	.66
Nonverbale und verbale Interaktionen	.71	.83	.63	.87
Emotionen und Gefühle	.46	.63	.49	.79
Sensorische Eindrücke	.58	.74	.45	.76
Gedächtnisprozesse und Memorieren	.18	.31	.19	.48
Gedanken	.45	.62	.44	.76
Räumliche Details	.37	.54	.40	.72
Zeitliche Details	.47	.64	.43	.75
Details	.37	.54	.42	.75
Persönliche Bedeutsamkeit	.37	.54	.43	.75
Klarheit und Lebendigkeit	.27	.43	.20	.50
Realismus und logische Struktur	.23	.38	.18	.47

Anm. N = 176. M(r) basiert für die freien Berichte auf den unabhängigen Beurteilungen derselben Aussagen durch zwei, für die Interviews durch vier Raterinnen. $r_{SP(2)}$ bezeichnet Spearman-Brown-korrigierte Werte für zwei, $r_{SP(4)}$ für vier Rater.

Für die Interviews variierten die mittleren Korrelationen zwischen r = .63 für verbale und nonverbale Interaktionen und r = .18 für Realismus und logische Struktur. Durch die Zusammenfassung von vier unabhängigen Beurteilungen ließ sich die Inter-Rater-Reliabilität substanziell verbessern. Für elf Merkmale ergaben sich zufrieden stellende Korrelationen von $r_{SP(4)} > .60$, von denen acht Merkmale sehr gute Inter-Rater-Reliabilitäten erzielten ($r_{SP(4)} \geq .75$). Für vier weitere Merkmale variierten die Korrelationen zwischen $.50 \leq r_{SP(4)} < .60$ und für die beiden übrigen Merkmale waren noch Korrelationen von $.47 \leq r_{SP(4)} < .50$ nachweisbar. Insgesamt wurden die Interviews in der vorliegenden Untersuchung ausreichend reliabel beurteilt.

Validität der ARJS-STV-G

Die Validität der ARJS wurde zunächst getrennt für die freien Berichte und die Interviews untersucht. Dazu wurden 2 x 2 MANOVAs mit dem Wahrheitsstatus und der Vorbereitungszeit als unabhängigen und den Beurteilungen der 17 ARJS-STV-G-Merkmale als abhängigen Variablen gerechnet. Danach erfolgte eine gemeinsame Analyse der Aussagequalität für die freien Berichte und der Interviews. Neben dem Wahrheitsstatus und der Vorbereitungzeit wurde die Aussageform im Rahmen einer 2 x 2 (x 2) MANOVA als Meßwiederholungsfaktor berücksichtigt. Um die Interpretation der Befunde zu erleichtern und deren Vergleichbarkeit mit anderen Untersuchungen zu ermöglichen, wurden zudem Effektstärkenmaße r und f abgeleitet. Schließlich wurde die Klassifikationsgüte der Merkmale diskriminanzanalytisch ermittelt.

Freie Berichte

Der Wahrheitsstatus hatte einen signifikanten Effekt auf die Beurteilung der 17 ARJS-STV-G-Merkmale für die freien Berichte, Wilks' Lambda = .83, $F(17,156) = 1.92$, $p = .020$, partielles $eta^2 = .17$. Die Mittelwertsunterschiede zwischen wahren und erfundenen freien Berichten sind in Tabelle 4.4 aufgeführt.

Tabelle 4.4

Mittelwertsunterschiede zwischen wahren und erfundenen Ereignissen hinsichtlich der 17 ARJS-STV-G-Merkmale für die freien Berichte (B01.01-B11.16)

Merkmale	$M_{erfunden}$	M_{wahr}	$F(1,172)$	p	r
Negative Äußerungen über sich selbst	1.60	2.03	6.57	**.011**	**.19**
Spontane Korrekturen	1.47	1.49	0.01	.914	.01
Erinnerungslücken zugeben	1.84	1.93	0.19	.667	.03
Komplikationen	2.06	2.30	1.54	.216	.09
Ungewöhnliche Details	1.84	2.35	7.32	**.008**	**.20**
Überflüssige Details	2.59	3.15	6.28	**.013**	**.19**
Nonverbale und verbale Interaktionen	3.78	4.31	3.73	.055	.15
Emotionen und Gefühle	4.59	5.10	6.85	**.010**	**.20**
Sensorische Eindrücke	1.80	1.97	0.58	.448	.06
Gedächtnisprozesse und Memorieren	3.40	3.76	2.16	.143	.11
Gedanken	4.19	4.82	5.96	**.016**	**.18**
Räumliche Details	3.26	3.27	0.01	.927	.01
Zeitliche Details	3.84	4.28	3.99	**.047**	**.15**
Details	3.75	4.34	12.73	**.000**	**.26**
Persönliche Bedeutsamkeit	5.06	5.15	0.19	.668	.03
Klarheit und Lebendigkeit	5.17	5.68	10.68	**.001**	**.24**
Realismus und logische Struktur	5.54	5.93	7.72	**.006**	**.21**

Anm. N = 176; 7-stufige Ratingskalen. Positive Effektstärken r signieren eine höhere Ausprägung in wahren Berichten.

Es zeigte sich für neun Merkmale eine signifikant höhere Aussagequalität bei wahren als bei erfundenen Aussagen. Die stärksten Effekte mittlerer Größenordnung zeigten sich für die Merkmale Details sowie Klarheit und Lebendigkeit. Für die Merkmale Realismus und logische Struktur, ungewöhnliche Details, Emotionen und Gefühle, negative Äußerungen über sich selbst, überflüssige Details und Gedanken waren Unterschiede mit mittleren bis geringen Effektstärken nachweisbar. Zudem zeigten sich geringe Effekte für die Beurteilung von zeitlichen Details und von nonverbalen und verbalen Interaktionen, wobei der Mittelwertsunterschied für das letztgenannte Merkmal marginal signifikant war. Ein geringer Effekt hinsichtlich des Merkmals Gedächtnisprozesse und Memorieren verfehlte statistische Signifikanz. Alle Merkmale waren erwartungsgemäß bei wahren Berichten stärker ausgeprägt als bei erfundenen.

Multivariat ließ sich kein signifikanter Haupteffekt der Vorbereitungszeit auf die 17 ARJS-STV-G-Merkmale nachweisen, Wilks' Lambda = .87, $F(17,156) = 1.40$, $p = .143$, partielles $eta^2 = .13$. Tabelle 4.5 ist jedoch zu entnehmen, dass sich auf univariater Ebene für acht Merkmale signifikante Effekte ergaben. Diese verwiesen auf eine höhere Aussagequalität für vorbereitete im Vergleich zu unvorbereiteten freien Berichten. Es resultierten Effektstärken mittlerer bis geringer Größenordnung für nonverbale und verbale Interaktionen, Details, Klarheit und Lebendigkeit, Realismus und logische Struktur, Komplikationen, zeitliche Details, überflüssige Details sowie für Gedächtnisprozesse und Memorieren. Zudem lag ein geringer Effekt für ungewöhnliche Details vor, wobei der Mittelwertsunterschied marginal signifikant war.

Die multivariate Wechselwirkung zwischen dem Wahrheitsstatus und der Vorbereitungszeit verfehlte ebenfalls statistische Signifikanz, Wilks' Lambda = .86, $F(17,156) = 1.54$, $p = .088$, partielles $eta^2 = .14$. Allerdings zeigte sich bei den univariaten Analysen eine signifikante Interaktion hinsichtlich des Merkmals Komplikationen, $F(1,172) = 9.18$, p .003, $f = .23$. Ohne Vorbereitung zeigten sich keine bedeutsamen Mittelwertsunterschiede zwischen wahren ($M = 1.75$) und

Tabelle 4.5

Mittelwertsunterschiede zwischen unvorbereiteten und vorbereiteten Aussagen hinsichtlich der 17 ARJS-STV-G-Merkmale für die freien Berichte (B01.01-B11.16)

Merkmale	$M_{unvorbereitet}$	$M_{vorbereitet}$	F(1,172)	p	r
Negative Äußerungen über sich selbst	1.88	1.76	0.49	.485	-.05
Spontane Korrekturen	1.39	1.57	1.26	.264	.08
Erinnerungslücken zugeben	1.74	2.02	1.68	.197	.10
Komplikationen	1.93	2.43	6.59	**.011**	**.19**
Ungewöhnliche Details	1.92	2.28	3.70	.056	.14
Überflüssige Details	2.62	3.12	4.86	**.029**	**.16**
Nonverbale und verbale Interaktionen	3.61	4.48	10.09	**.002**	**.23**
Emotionen und Gefühle	4.74	4.94	1.06	.305	.08
Sensorische Eindrücke	1.80	1.97	0.58	.448	.06
Gedächtnisprozesse und Memorieren	3.32	3.84	4.55	**.034**	**.16**
Gedanken	4.30	4.71	2.46	.119	.12
Räumliche Details	3.10	3.43	3.02	.084	.13
Zeitliche Details	3.80	4.32	5.29	**.023**	**.17**
Details	3.82	4.27	7.53	**.007**	**.20**
Persönliche Bedeutsamkeit	5.00	5.21	1.13	.290	.08
Klarheit und Lebendigkeit	5.23	5.62	6.42	**.012**	**.19**
Realismus und logische Struktur	5.55	5.92	6.84	**.010**	**.19**

Anm. N = 176; 7-stufige Ratingskalen. Positive Effektstärken r signieren eine höhere Ausprägung in vorbereiteten Berichten.

erfundenen Berichten (M = 2.10), F(1,172) = 1.60, p = .208, r = .10. Hingegen enthielten wahre Berichte (M = 2.85) mehr Hinweise auf Komplikationen als erfundene (M = 2.01), wenn Gelegenheit zur Vorbereitung bestand, F(1,172) = 9.11, p = .003, r = .22.[1]

Schließlich wurde eine multiple Diskriminanzanalyse durchgeführt, um die Klassifikationsgüte der ARJS-STV-G zu bestimmen. Dabei wurden die 17 Glaubhaftigkeitsmerkmale als Prädiktorvariablen und der objektive Wahrheitsstatus der 176 freien Berichte als Klassifizierungsvariable verwendet. Die resultierende Diskriminanzfunktion erwies sich als signifikant, Wilks' Lambda

---

[1] Die im vorangegangenen Experiment durchgeführten Analysen des Stimulusmaterials zeigten, dass wahre Aussagen umfangreicher waren als erfundene. Ebenso erwiesen sich vorbereitete im Vergleich zu unvorbereiteten Aussagen als länger. Obwohl anzunehmen ist, dass Skalenbeurteilungen relativ robust gegenüber Variationen in der Aussagelänge sind (vgl. Granhag et al., 2006, Strömwall et al., 2004), wurde zusätzlich eine 2 x 2 MANCOVA durchgeführt, bei der die Anzahl der Wörter als Kovariate verwendet wurde. Der Wahrheitsstatus und die Vorbereitungszeit dienten als unabhängige Faktoren und die 17 ARJS-STV-G-Merkmale als abhängige Variablen. Die Ergebnisse finden sich in Anhang 4c. Für die freien Berichte war bei Kontrolle der Anzahl an Wörtern der multivariate Effekt des Wahrheitsstatus auf die inhaltlichen Aussagemerkmale nicht mehr nachweisbar, Wilks' Lambda = .86, F(17,155) = 1.44, p = .125, partielles eta^2 = .14. Univariat zeigten sich jedoch noch signifikante Mittelwertsunterschiede für die Merkmale Details (r = .18), Klarheit und Lebendigkeit (r = .17) sowie negative Äußerungen über sich selbst (r = .15). Zudem ergab die Kovarianzanalyse in Übereinstimmung mit den dargestellten Befunden keinen multivariaten Haupteffekt der Vorbereitungszeit, Wilks' Lambda = .93, F(17,155) = 0.72, p = .776, partielles eta^2 = .07, und keine Wechselwirkung mit dem Wahrheitsstatus, Wilks' Lambda = .86, F(17,155) = 1.52, p = .094, partielles eta^2 = .14.

= .83, $Chi^2(17) = 30.61$, $p = .022$, und erlaubte es 66.5% ($n = 117$) der 176 Berichte korrekt zuzuordnen. Von den 88 wahren Aussagen wurden 63.6% ($n = 56$), von den 88 erfundenen 69.3% ($n = 61$) richtig klassifiziert.

Interviews

Für die Interviews ergab sich ebenfalls ein signifikanter Effekt des Wahrheitsstatus auf die ARJS-STV-G-Merkmale, Wilks' Lambda = .74, $F(17,156) = 3.18$, $p < .001$, partielles $eta^2 = .26$. In Tabelle 4.6 sind die Ergebnisse der univariaten Analysen dargestellt. Mit Ausnahme von sensorischen Eindrücken waren die Mittelwerte jeweils höher in wahren als erfundenen Aussagen. Es zeigte sich für zehn Merkmale eine signifikant höhere Qualität von wahren im Vergleich zu erfundenen Aussagen. Starke bis mittlere Effekte waren für die Merkmale Realismus und logische Struktur, überflüssige Details, Details, nonverbale und verbale Interaktionen sowie Klarheit und Lebendigkeit nachweisbar. Für ungewöhnliche Details resultierte ein Effekt mittlerer Größenordnung. Mittlere bis geringe Effekte zeigten sich zudem für Gedanken, Emotionen und Gefühle, sowie für Komplikationen. Für spontane Korrekturen und Gedächtnisprozesse ergaben sich ebenfalls signifikante bzw. marginal signifikante Mittelwertsunterschiede. Des Weiteren wurden Effekte geringer Größenordnung für negative Äußerungen über sich selbst, räumliche und zeitliche Details aufgedeckt, die jedoch keine statistische Signifikanz erzielten.

Für die Interviews zeigte sich kein Haupteffekt der Vorbereitungszeit, weder auf multivariater, Wilks' Lambda = .94, $F(17,156) = 0.64$, $p = .850$, partielles $eta^2 = .07$, noch auf univariater Ebene, alle $Fs(1,172) \leq 3.19$, $ps \geq .076$, $rs \leq .14$. Ebenso lag keine multivariate Wechselwirkung zwischen dem Wahrheitsstatus und der Vorbereitungszeit vor, Wilks' Lambda = .91, $F(17,156) = 0.94$, $p = .530$, partielles $eta^2 = .09$. Allerdings war auf univariater Ebene eine Wechselwirkung der beiden Faktoren hinsichtlich der persönlichen Signifikanz erkennbar. Bei vorbereiteten Interviews ergaben sich keine Mittelwertsunterschiede zwischen wahren ($M = 5.02$) und erfundenen Aussagen ($M = 5.25$), $F(1,172) = 1.09$, $p = .298$, $r = .08$. Hingegen

Tabelle 4.6

Mittelwertsunterschiede zwischen wahren und erfundenen Ereignissen hinsichtlich der 17 ARJS-STV-G-Merkmale für die Interviews (C01.01-C11.16)

Merkmale	$M_{erfunden}$	M_{wahr}	$F(1,172)$	p	r
Negative Äußerungen über sich selbst	1.70	1.96	2.77	.098	.13
Spontane Korrekturen	1.49	1.76	3.89	**.050**	**.15**
Erinnerungslücken zugeben	3.04	3.19	0.34	.563	.04
Komplikationen	1.98	2.35	5.24	**.023**	**.17**
Ungewöhnliche Details	1.83	2.34	12.18	**.001**	**.26**
Überflüssige Details	2.85	3.53	19.17	**.000**	**.32**
Nonverbale und verbale Interaktionen	3.55	4.47	14.69	**.000**	**.28**
Emotionen und Gefühle	4.61	5.05	5.65	**.019**	**.18**
Sensorische Eindrücke	2.05	2.02	0.02	.883	-.01
Gedächtnisprozesse und Memorieren	3.60	3.91	3.84	.052	.15
Gedanken	4.38	4.84	6.21	**.014**	**.19**
Räumliche Details	4.05	4.30	2.14	.145	.11
Zeitliche Details	5.09	5.30	1.69	.195	.10
Details	4.17	4.67	15.78	**.000**	**.29**
Persönliche Bedeutsamkeit	5.02	5.15	0.66	.419	.06
Klarheit und Lebendigkeit	5.41	5.85	14.12	**.000**	**.28**
Realismus und logische Struktur	5.68	6.13	22.04	**.000**	**.34**

Anm. N = 176; 7-stufige Ratingskalen. Positive Effektstärken r signieren eine höhere Ausprägung in wahren Interviews.

wurden unvorbereitete wahre Interviews (M = 5.28) als persönlich bedeutsamer beurteilt als unvorbereitete erfundene (M = 4.80), F(1,172) = 4.80, p = .030, r = .16.[2]

Eine Diskriminanzanalyse erlaubte eine signifikante Vorhersage des tatsächlichen Wahrheitsstatus anhand der 17 ARJS-STV-G-Merkmale, Wilks' Lambda = .743, Chi^2(17) = 49.08, p < .001. Von den insgesamt 176 Interviews ließen sich 72.7% (n = 128) korrekt klassifizieren. Dabei wurde für die 88 wahren Aussagen (76.1%, n = 67) eine bessere Klassifikationsgüte erzielt als für die 88 erfundenen (69.3%, n = 61).

---

[2] Auch für die Interviews zeigte sich, dass wahre Aussagen umfangreicher waren als erfundene. Daher wurde erneut eine multiple Kovarianzanalyse, bei der die Anzahl der Wörter kontrolliert wurde, gerechnet. Deren univariate Ergebnisse sind im Anhang 4d aufgeführt. Die Befunde der 2 x 2 MANCOVA und der 2 x 2 MANOVA waren weitestgehend vergleichbar. Es zeigte sich ein signifikanter Haupteffekt des Wahrheitsstatus auf die ARJS-STV-G-Merkmale, Wilks' Lambda = .75, F(17,155) = 2.99, p < .001, partielles eta^2 = .25. Erneut ergab sich eine signifikant höhere Aussagequalität wahrer im Vergleich zu erfundenen Aussagen für die Merkmale Realismus und logische Struktur (r = .30), überflüssige Details (r = .27), Details (r = .24), nonverbale und verbale Interaktionen (r = .23), ungewöhnliche Details (r = .22) sowie Klarheit und Lebendigkeit (r = .22). Die Mittelwertsunterschiede hinsichtlich der Merkmale Komplikationen, spontane Korrekturen, Gedanken sowie Emotionen und Gefühle erzielten hingegen nicht mehr statistische Signifikanz. Das Ausmaß an Vorbereitungszeit wirkte sich nicht auf die Beurteilung der ARJS-STV-G-Merkmale aus, Wilks' Lambda = .94, F(17,155) = 0.63, p = .862, partielles eta^2 = .07. Ebenso war keine Wechselwirkung zwischen dem Wahrheitsstatus und der Vorbereitungszeit erkennbar, Wilks' Lambda = .90, F(17,155) = 1.04, p = .421, partielles eta^2 = .10.

Vergleich der ARJS-STV-G-Beurteilungen beider Aussageformen

Um zu überprüfen, ob die Beurteilungen der Aussagequalität mit der Aussageform variierten, wurde eine weitere 2 x 2 (x 2) MANOVA durchgeführt. Erneut wurden die 17 ARJS-STV-G-Merkmale als abhängige Variablen und der Wahrheitsstatus und die Vorbereitungszeit als unabhängige Between-Subjects-Faktoren verwendet. Zusätzlich wurde die Aussageform (freie Berichte versus Interviews) als Within-Subjects-Faktor berücksichtigt.

Hinsichtlich der Between-Subjects-Faktoren Wahrheitsstatus und Vorbereitungszeit wurden die Befunde der getrennten Analysen für die freien Berichte und die Interviews repliziert. Bei der gemeinsamen Analyse beider Aussageformen variierten die Beurteilungen der Aussagequalität erneut in Abhängigkeit vom Wahrheitsstatus, Wilks' Lambda = .78, $F(17,156) = 2.62$, $p = .001$, partielles $eta^2 = .22$. Wahre Aussagen erzielten signifikant höhere Beurteilungen als erfundene für zehn Merkmale. Für Details ($r = .30$), Realismus und logische Struktur ($r = .30$), überflüssige Details ($r = .29$) sowie Klarheit und Lebendigkeit ($r = .29$) ergaben sich starke bis mittlere Effekte. Für ungewöhnliche Details ($r = .26$), nonverbale und verbale Interaktionen ($r = .23$), Gedanken ($r = .21$), Emotionen und Gefühle ($r = .21$), negative Äußerungen über sich selbst ($r = .18$) sowie Gedächtnisprozesse und Memorieren ($r = .15$) zeigten sich Effektstärken mittlerer bis geringer Größenordnung. Des Weiteren verfehlten die erwartungsgemäßen Mittelwertsunterschiede zwischen wahren und erfundenen Aussagen hinsichtlich der beiden Merkmale zeitliche Details ($F(1,172) = 3.71$, $p = .056$, $r = .15$) und Komplikationen ($F(1,172) = 3.48$, $p = .064$, $r = .14$) nur knapp statistische Signifikanz. Ein Haupteffekt der Vorbereitungszeit war über beide Aussagen hinweg nicht nachweisbar, Wilks' Lambda = .91, $F(17,156) = 0.90$, $p = .579$, partielles $eta^2 = .09$. Zudem lag erneut keine multivariate Wechselwirkung zwischen dem Wahrheitsstatus und der Vorbereitungszeit vor, Wilks' Lambda = .87, $F(17,156) = 1.32$, $p = .188$, partielles $eta^2 = .13$.

Die Aussageform hatte einen signifikanten Effekt auf die Beurteilung der ARJS-STV-G-Merkmale, Wilks' Lambda = .33, $F(17,156) = 18.58$, $p < .001$, partielles $eta^2 = .67$. Tabelle 4.7 ist zu entnehmen, dass sich für sieben Skalen signifikante Mittelwertsunterschiede zeigten. Die Aussagequalität der Interviews wurde höher eingeschätzt als die der freien Berichte. Starke Effekte ergaben sich für die Merkmale zeitliche Details, räumliche Details und Erinnerungslücken zugeben. Effekte mittlerer bis geringer Größenordnung resultierten für das Merkmal Details. Schließlich zeigten sich geringe Effekte für überflüssige Details, Klarheit und Lebendigkeit, Realismus und logische Struktur.

Eine multivariate Interaktion der Aussageform und des Wahrheitsstatus war nicht feststellbar, Wilks' Lambda = .88, $F(17,156) = 1.25$, $p = .234$, partielles $eta^2 = .12$. Auch auf univariater Ebene waren keine signifikanten Effekte erkennbar, alle Fs$(1,172) \leq 3.16$, ps $\geq .077$, fs $\leq .14$. Im Gegensatz dazu interagierte die Aussageform mit der Vorbereitungszeit, Wilks' Lambda = .81, $F(17,156) = 0.81$, $p = .009$, partielles $eta^2 = .19$. Signifikante Interaktionseffekte ergaben sich für Komplikationen ($f = .34$), Details ($f = .21$), ungewöhnliche Details ($f = .17$), zeitliche Details ($f = .17$), Gedächtnisprozesse und Memorieren ($f = .17$) sowie für nonverbale und verbale Interaktionen ($f = .15$). Dabei erzielten vorbereitete freie Berichte höhere Bewertungen als unvorbereitete (vgl. Tabelle 4.5), während für die Interviews keine Mittelwertsunterschiede in Abhängigkeit von der Vorbereitungszeit festzustellen waren. Es lag keine signifikante Wechselwirkung zwischen den drei Faktoren Aussageform, Wahrheitsstatus und Vorbereitung vor, Wilk's Lambda = 0.91, $F(17,156) = 0.95$, $p = .513$, partielles $eta^2 = .09$.

Tabelle 4.7

Mittelwertsunterschiede zwischen freien Berichten und Interviews hinsichtlich der 17 ARJS-STV-G Merkmale

Merkmale	M_{FB}	$M_{Interv.}$	$F(1,172)$	p	r
Negative Äußerungen über sich selbst	1.82	1.83	0.02	.891	.00
Spontane Korrekturen	1.48	1.62	2.59	.109	.07
Erinnerungslücken zugeben	1.88	3.12	92.58	**.000**	**.36**
Komplikationen	2.18	2.16	0.04	.846	-.01
Ungewöhnliche Details	2.10	2.08	0.02	.879	-.01
Überflüssige Details	2.87	3.19	7.78	**.006**	**.12**
Nonverbale und verbale Interaktionen	4.04	4.01	0.09	.761	-.01
Emotionen und Gefühle	4.84	4.83	0.02	.891	-.01
Sensorische Eindrücke	1.88	2.03	2.55	.112	.06
Gedächtnisprozesse und Memorieren	3.58	3.75	2.15	.145	.06
Gedanken	4.51	4.61	0.84	.361	.03
Räumliche Details	3.26	4.17	93.64	**.000**	**.35**
Zeitliche Details	4.06	5.20	141.75	**.000**	**.40**
Details	4.05	4.42	32.47	**.000**	**.18**
Persönliche Bedeutsamkeit	5.11	5.09	0.09	.763	-.01
Klarheit und Lebendigkeit	5.42	5.63	9.53	**.002**	**.11**
Realismus und logische Struktur	5.73	5.91	8.21	**.005**	**.10**

Anm. 7-stufige Ratingskalen. Positive Effektstärken r signieren eine höhere Ausprägung in Interviews.

Analysen zur Valenz der Ereignisse

Im Rahmen des vorangegangen Experiments wurde die Valenz des verwendeten Stimulusmaterials durch vier trainierte Raterinnen eingeschätzt. Die Beurteilung erfolgte anhand siebenstufiger Ratingskalen. Dabei verwiesen geringe Werte auf eine negative und hohe auf eine positive Erlebnisqualität der geschilderten Ereignisse. Wie im vorangegangen Experiment beschrieben, wurden die Urteile der vier Raterinnen zunächst gemittelt und dann am Median dichotomisiert. Anhand dieser Variable wurde untersucht, ob sich die Valenz der Ereignisse auf die Beurteilungen der 17 ARJS-STV-G-Merkmale auswirkt. Um potenzielle Wechselwirkungseffekte aufzudecken, ging zudem der Wahrheitsstatus in die Analysen mit ein. Es wurden getrennte 2 x 2 MANOVAs für die freien Berichte und die Interviews berechnet.

Freie Berichte

Für die freien Berichte war kein multivariater Haupteffekt der Valenz der geschilderten Ereignisse auf die Beurteilung der Aussagequalität festzustellen, Wilks' Lambda = .87, $F(17,156) = 1.40$, $p = .143$, partielles $eta^2 = .13$. Zudem zeigte sich keine signifikante Wechselwirkung zwischen dem Wahrheitsstatus und der Valenz, Wilks' Lambda = .91, $F(17,156) = 0.88$, $p = .594$, partielles $eta^2 = .08$, 1-beta = .60.

Interviews

Im Gegensatz zu den freien Berichten ergab sich für die Interviews ein marginal signifikanter Haupteffekt der Valenz, Wilks' Lambda = .85, $F(17,156) = 1.65$, $p = .058$, partielles $eta^2 = .15$. Tabelle 4.8 ist zu entnehmen, dass sich dieser auf signifikante Mittelwertsunterschiede für sechs Merkmale zurückführen ließ.

Tabelle 4.8

Mittelwertsunterschiede zwischen positiv und negativ eingestuften Ereignissen hinsichtlich der 17 ARJS-STV-G-Merkmale für die Interviews (C01.01-C11.16)

Merkmale	$M_{negativ}$	$M_{positiv}$	F(1,172)	p	r
Negative Äußerungen über sich selbst	1.85	1.80	0.12	.730	-.03
Spontane Korrekturen	1.58	1.67	0.42	.518	.05
Erinnerungslücken zugeben	2.95	3.28	1.63	.204	.10
Komplikationen	2.07	2.26	1.27	.261	.09
Ungewöhnliche Details	2.09	2.08	0.01	.923	-.01
Überflüssige Details	3.23	3.15	0.24	.627	-.04
Nonverbale und verbale Interaktionen	4.32	3.70	**6.91**	**.009**	**-.20**
Emotionen und Gefühle	5.07	4.59	**6.96**	**.009**	**-.20**
Sensorische Eindrücke	1.99	2.08	0.22	.637	.04
Gedächtnisprozesse und Memorieren	3.97	3.53	**7.89**	**.006**	**-.21**
Gedanken	4.84	4.38	**6.31**	**.013**	**-.19**
Räumliche Details	4.07	4.27	1.33	.250	.09
Zeitliche Details	5.07	5.33	2.58	.110	.12
Details	4.50	4.34	1.58	.210	-.10
Persönliche Bedeutsamkeit	5.25	4.93	**4.18**	**.042**	**-.15**
Klarheit und Lebendigkeit	5.76	5.50	**5.02**	**.026**	**-.17**
Realismus und logische Struktur	5.97	5.85	1.56	.213	-.09

Anm. N = 176; 7-stufige Ratingskalen. Positive Effektstärken r signieren eine höhere Ausprägung in positiv eingestuften Ereignissen.

Dabei erzielten negative Ereignisse eine höhere Aussagequalität als positive. Es ergaben sich Effekte mittlerer bis geringer Größenordnung für Gedächtnisprozesse und Memorieren (r = -.21), Emotionen und Gefühle (r = -.20), nonverbale und verbale Interaktionen (r = -.20), Gedanken (r = -.19), Klarheit und Lebendigkeit (r = -.17) sowie für die persönliche Bedeutsamkeit (r = -.15).

Zudem erzielte die multivariate Wechselwirkung zwischen dem Wahrheitsstatus und der Valenz der geschilderten Ereignisse statistische Signifikanz, Wilks' Lambda = .81, $F(17,156) = 2.22$, p = .006, partielles eta^2 = .20, 1-beta = .98. Dies ließ sich auf die drei Merkmale Gedanken (f = .18), räumliche (f = .17) und zeitliche Details (f = .15) zurückführen. Für negative Ereignisse zeigten sich keine Mittelwertsunterschiede hinsichtlich des Merkmals Gedanken bei wahren (M = 4.86) und bei erfundenen Interviews (M = 4.81), $F(1,172) = 0.04$, p = .842, r = .02. Für positive Ereignisse hingegen umfassten wahre Aussagen (M = 4.82) mehr Hinweise auf Gedanken als erfundene (M = 3.94), $F(1,172) = 11.83$, p = .001, r = .25. Hinsichtlich des Merkmals räumliche Details ergaben sich ähnliche Unterschiede. Wenn negative Ereignisse geschildert wurden, war das Ausmaß an räumlichen Details für wahre (M = 4.03) und erfundene Aussagen (M = 4.11) vergleichbar, $F(1,172) = 0.11$, p = .744, r = .02. Bei positiven Ereignissen hingegen erzielten wahre Aussagen (M = 4.57) höhere Beurteilungen als erfundene (M = 3.98), $F(1,172) = 5.88$, p = .016, r = .18. Für das Merkmal zeitliche Details zeigten sich allerdings gegensätzliche Befunde. Für negative Ereignisse ließen sich signifikante Mittelwertsunterschiede zwischen wahren (M = 5.35) und erfundenen Interviews (M = 4.78) nachweisen, $F(1,172) = 6.23$, p = .014, r = .19. Hingegen war für positive Ereignisse die Aussagequalität wahrer (M = 5.26) und erfundener Interviews (M = 5.40) vergleichbar, $F(1,172) = 0.39$, p = .533, r = .05.

Diskussion

Die vorliegende Untersuchung zielte darauf ab, die Reliabilität und Validität der ARJS-STV-G, einer Kurzform der ARJS, zu untersuchen. Des Weiteren wurden

der Einfluss der Gelegenheit zur Vorbereitung, der Aussageform sowie der Valenz des geschilderten Ereignisses auf die ARJS-STV-G-Beurteilungen untersucht. Dazu wurde das bereits im vorangegangenen Experiment verwendete Stimulusmaterial erneut von acht unabhängigen Raterinnen anhand der ARJS-STV-G beurteilt.

Reliabilität der ARJS-STV-G

Die Raterinnen verfügten über keinerlei Vorwissen hinsichtlich der inhaltlichen Glaubhaftigkeitsanalyse und wurden nur kurz mit der Anwendung der ARJS-STV-G vertraut gemacht. Daher überrascht es kaum, dass sich für die Einzelratings keine guten Inter-Rater-Reliabilitäten zeigten. Lediglich für das ARJS-STV-G-Merkmal nonverbale und verbale Interaktionen war eine hohe Konvergenz der Beurteilungen festzustellen.

Die Inter-Rater-Reliabilitäten wurden jedoch durch die Zusammenfassung der Beurteilungen von mehreren Raterinnen substanziell verbessert. Bei den freien Berichten konnten Spearman-Brown-Korrekturen für zwei, bei den Interviews für vier Rater vorgenommen werden. Insbesondere für die Interviews wurden dadurch oftmals sehr gute Inter-Rater-Reliabilitäten erzielt. Doch auch für die freien Berichte sind die meisten Werte tolerierbar. Lediglich die Skalen Realismus und logische Struktur, ungewöhnliche Details, Gedächtnisprozesse und Memorieren sowie spontane Korrekturen wurden nicht reliabel erfasst. Wahrscheinlich haben auch Boden- und Deckeneffekte dazu beigetragen. So waren für die Skalen spontane Korrekturen und ungewöhnliche Details deutlich reduzierte bzw. für Realismus und logische Struktur erhöhte Mittelwerte festzustellen.

Insgesamt ist zu schlussfolgern, dass sich die ARJS-STV-G-Merkmale reliabel erfassen lassen, wenn mehrere Rater für die Aussageanalyse eingesetzt werden. Aufgrund der vorliegenden Befunde ist anzunehmen, dass bereits bei vier Ratern eine zufrieden stellende Reliabilität gewährleistet werden kann.

Validität der ARJS-STV-G

Es wurde angenommen, dass wahre Aussagen höhere Beurteilungen hinsichtlich der 17 ARJS-STV-G-Merkmale erhalten als erfundene. Diese Hypothese wurde durch die vorliegenden Untersuchungsbefunde unterstützt. Sowohl für die freien Berichte als auch für die Interviews zeigte sich ein signifikanter Haupteffekt des Wahrheitsstatus auf die Aussagequalität. Bei den freien Berichten ergaben sich für neun, bei den Interviews für zehn ARJS-STV-G-Merkmale erwartungsgemäße Mittelwertsunterschiede.

Einschränkend ist jedoch darauf hinzuweisen, dass für die freien Berichte kein multivariater Effekt des Wahrheitsstatus mehr nachweisbar war, wenn die Anzahl der Wörter kontrolliert wurde. Bei dem verwendeten Stimulusmaterial waren wahre Aussagen länger als erfundene, und längere Aussagen wiesen eine höhere inhaltliche Qualität auf als kürzere. Die Raterinnen berücksichtigten bei ihren Skalenbeurteilungen offenbar nur unzureichend, dass umfangreichere Aussagen zwangsläufig mehr Informationen enthalten. Die höhere Qualität wahrer im Vergleich zu erfundenen freien Berichten schien also vor allem auf die Länge der Aussage zurückzuführen zu sein. Für die Interviews war hingegen auch dann ein multivariater Effekt des Wahrheitsstatus auf die Beurteilung der ARJS-STV-G-Merkmale nachweisbar, wenn die Anzahl der Wörter als Kovariate berücksichtigt wurde.

Da sich keine multivariate Wechselwirkung zwischen dem Wahrheitsstatus und der Aussageform zeigte, wurden beide Aussageformen zusätzlich gemeinsam analysiert. Für zehn der 17 ARJS-STV-G-Merkmale ergaben sich signifikante, für zwei weitere marginal signifikante Mittelwertsunterschiede. Alle Merkmale waren erwartungsgemäß bei wahren Aussagen stärker ausgeprägt als bei erfundenen. Für die Skalen Realismus und logische Struktur, Details, Klarheit und Lebendigkeit, überflüssige Details sowie ungewöhnliche Details ergaben sich starke bis mittlere Effekte. Für nonverbale und verbale Interaktionen, Gedanken, Emotionen und Gefühle, negative Äußerungen über sich selbst sowie

für Gedächtnisprozesse und Memorieren ergaben sich Effekte mittlerer bis geringer Größenordnung. Die Schilderung von Gedanken war im Gegensatz zu den Vorhersagen des Realitätsüberwachungs-Ansatzes (Johnson & Raye, 1981) erneut eher bei wahren als bei erfundenen Aussagen vorzufinden. Dies unterstützt die im Rahmen der ARJS formulierte Annahme, dass Gedanken als Glaubhaftigkeitsmerkmal aufzufassen sind. Schließlich waren die Mittelwertsunterschiede hinsichtlich der Skalen zeitliche Details und Komplikationen marginal signifikant und resultierten in geringen Effektstärken. Demnach haben sich sowohl ARJS-STV-G-Merkmale mit forensischem (z.B. nonverbale und verbale Interaktionen), sozialpsychologischem (z.B. negative Äußerungen über sich selbst) und gedächtnispsychologischem (z.B. Klarheit und Lebendigkeit) Entwicklungshintergrund als valide erwiesen.

Die für die Kurzform ermittelten Effektstärken sind in Tabelle 4.9 den Ergebnissen der vorangegangenen Untersuchung zur Validität der Langform vergleichend gegenübergestellt. Bei der Langform zeigten sich Mittelwertsunterschiede zwischen wahren und erfundenen Aussagen für die Skala Fehler und sozial Unerwünschtes. Auch für das ARJS-STV-G-Merkmal negative Äußerungen über sich selbst lagen erwartungsgemäße Mittelwertsunterschiede vor. Die beiden ebenfalls unter der Skala Fehler und sozial Unerwünschtes einzuordnenden ARJS-STV-G-Merkmale spontane Korrekturen und Erinnerungslücken zugeben bildeten hingegen keine Qualitätsunterschiede in Abhängigkeit vom Wahrheitsstatus ab. Im Gegensatz dazu konnte für die Skala Realismus und logische Struktur der Langform aufgrund eines Deckeneffektes keine Validität nachgewiesen werden. Das entsprechende Merkmal der Kurzform differenzierte jedoch erwartungsgemäß zwischen wahren und erfundenen Aussagen. Für die übrigen Merkmale waren die Befunde der Lang- und Kurzform weitestgehend vergleichbar. Lediglich die Größenordnung der ermittelten Effektstärken fiel unterschiedlich aus. Dabei zeigte sich teilweise eine Überlegenheit für die vollständige ARJS-Analyse, teilweise für die Kurzform.

Tabelle 4.9

Validität der Kurz- und Langform der ARJS im Vergleich (Effektstärkemaße r)

Merkmale	ARJS-STV-G	ARJS
Negative Äußerungen über sich selbst	**.18**	**.16a**
Spontane Korrekturen	.09	**.16a**
Erinnerungslücken zugeben	.05	**.16a**
Komplikationen	**.14**	**.21b**
Ungewöhnliche Details	**.26**	**.21b**
Überflüssige Details	**.29**	**.32c**
Nonverbale und verbale Interaktionen	**.23**	**.16**
Emotionen und Gefühle	**.21**	**.25**
Sensorische Eindrücke	.03	-.02
Gedächtnisprozesse und Memorieren	**.15**	**.18**
Gedanken	**.21**	**.15**
Räumliche Details	.07	.08
Zeitliche Details	**.15**	**.15**
Details	**.30**	**.32c**
Persönliche Bedeutsamkeit	.05	-.11
Klarheit und Lebendigkeit	**.29**	**.31**
Realismus und logische Struktur	**.30**	.08

Anm. a = Effektstärken für die ARJS-Skala Fehler und sozial Unerwünschtes; b = ARJS-Skala Komplikationen und ungewöhnliche Details; c = ARJS-Skala Details. Effektstärken $r \geq .14$ sind fett gedruckt.

Insgesamt bleiben die Effektstärken für die ARJS-STV-G-Merkmale kaum hinter denen der Langform zurück. Allerdings ist darauf hinzuweisen, dass Sporer, Masip und Cramer (in Vorbereitung; siehe auch Cramer, 2005) weniger ermutigende Befunde für die spanischsprachige Kurzform der ARJS berichteten. Die Autoren fanden nur für die beiden Merkmale negative Äußerungen über sich selbst und Erinnerungslücken höhere Mittelwerte bei wahren als bei erfundenen Aussagen.

Die divergierenden Befunde der vorliegenden und der von Sporer et al. (in Vorbereitung) durchgeführten Studie lassen sich vermutlich auf Unterschiede im Stimulusmaterial und Untersuchungsdesign zurückführen. Das von Sporer et al. verwendete Stimulusmaterial umfasste 64 Aussagen zu moralischen und gesetzlichen Vergehen. Die Probanden formulierten im Rahmen von Interviews sowohl wahre als auch erfundene Aussagen, es wurde also eine Within-Subjects-Variation des objektiven Wahrheitsstatus vorgenommen. Allerdings war auch bei den erfundenen Aussagen ein Erlebnisbezug vorhanden, da die Tathandlung lediglich auf eine unschuldige Person übertragen wurde. Die Aussagen wurden von 32 Ratern, die jeweils 16 Aussagen bearbeiteten, anhand der ARJS-Kurzform beurteilt. Im Gegensatz dazu wurden in der vorliegenden Untersuchung wesentlich mehr Aussagen analysiert, die zudem umfangreicher waren. Dabei schilderte jede Probandin entweder ein wahres oder ein erfundenes Ereignis, es erfolgte also eine Between-Subjects-Variation des Wahrheitsstatus. Die ARJS-STV-G-Beurteilungen wurden durch acht Rater vorgenommen, die jeweils 48 Aussagen analysierten. Zudem wurden die Rater unterschiedlich auf die inhaltliche Aussageanalyse vorbereitet. Sporer et al. informierten die Rater ausschließlich schriftlich über die Definition der 17 Glaubhaftigkeitsmerkmale. Im Gegensatz dazu nahmen die Raterinnen in der vorliegenden Untersuchung an einer 2,5-stündigen Informationsveranstaltung teil. Dabei wurden inhaltliche Fragen geklärt, und es gab Gelegenheit, die Anwendung der ARJS-STV-G anhand von

Übungsmaterialien zu erproben. Zusätzlich erhielten sie eine deutschsprachige Fassung des von Sporer et al. verwendeten Informationsmaterials.

Möglicherweise reichten die kurzen schriftliche Informationen nicht aus, um die Rater auf die inhaltliche Aussageanalyse anhand der ARJS-Kurzform vorzubereiten. Eine kurze Informationsveranstaltung genügte jedoch, um für das in der vorliegenden Studie verwendete Stimulusmaterial eine valide Einschätzung der Aussagequalität anhand der ARJS-STV-G zu gewährleisten. Dennoch bleibt in weiteren Untersuchungen zu klären, ob die Befunde der vorliegenden Studie repliziert werden können.

Effekte der Vorbereitungszeit

Bevor die Probanden ihre freien Berichte abgaben, wurde die Gelegenheit zur Vorbereitung experimentell manipuliert. Es wurde erwartet, dass Aussagen nach einer 15-minütigen Vorbereitungszeit höhere ARJS-STV-G-Beurteilungen erzielen würden als unvorbereitete. Diese Hypothese ließ sich teilweise unterstützen. Für die freien Berichte wurde zwar kein multivariater Haupteffekt der Vorbereitungszeit auf die ARJS-STV-G-Merkmale nachgewiesen, bei den univariaten Analysen ergaben sich jedoch für acht Merkmale signifikante Mittelwertsunterschiede. Diese bildeten erwartungsgemäß eine höhere Qualität vorbereiteter Aussagen im Vergleich zu unvorbereiteten ab und manifestierten sich in Effektstärken mittlerer bis geringer Größenordnung. Im Gegensatz zu den freien Berichten waren für die Interviews keinerlei Unterschiede zwischen vorbereiteten und unvorbereiteten Aussagen festzustellen. Dies ist vermutlich darauf zurückzuführen, dass die experimentelle Manipulation eine Woche zuvor induziert worden war. Zum Zeitpunkt der Interviews waren die ursprünglich qualitätssteigernden Effekte der Vorbereitung nicht mehr wirksam.

Von übergeordnetem Interesse ist jedoch die Frage, ob die Gelegenheit zur Vorbereitung die Validität der ARJS-STV-G-Merkmale reduziert. Dies würde sich in einer signifikanten Wechselwirkung zwischen dem Wahrheitsstatus und der Vorbereitungszeit auf die Aussagequalität widerspiegeln. Multivariat ergab sich

weder für die freien Berichte noch für die Interviews eine Wechselwirkung. Allerdings waren für beide Aussageformen auf univariater Ebene Interaktionseffekte festzustellen. Bei den freien Berichten ergab sich ein signifikanter Mittelwertsunterschied für das ARJS-STV-G-Merkmal Komplikationen. Ohne die Gelegenheit zur Vorbereitung zeigten sich keine Mittelwertsunterschiede zwischen wahren und erfundenen Berichten. Wahre Berichte enthielten jedoch dann mehr Verweise auf Komplikationen als erfundene, wenn sie vorbereitet wurden.

Allerdings erscheint dieser Befund vor dem Hintergrund der bisherigen Forschung strittig. Vielmehr ist anzunehmen, dass insbesondere Personen, die Erfundenes berichten, die Gelegenheit zur Vorbereitung nutzen, um ihre Aussage überzeugend zu gestalten. Infolgedessen sollte sich die Vorbereitung einer Aussage auf inhaltliche Aussagemerkmale validitätsmindernd auswirken. Entsprechende Effekte wurden beispielsweise von Alonso-Quecuty (1992) sowie Sporer (1998) berichtet. Andererseits liegen auch Forschungsbefunde vor, nach denen die Validität inhaltlicher Aussagemerkmale nicht durch die Gelegenheit zur Vorbereitung beeinflusst (Sporer & Burghardt, 2004) oder sogar verbessert wird (Sporer 1997). Allerdings wurden für Merkmale, die dem ARJS-STV-G-Merkmal Komplikationen ähneln, bislang keine Wechselwirkungen zwischen dem Wahrheitsstatus und der Gelegenheit zur Vorbereitung festgestellt (Sporer, 1997; Sporer & Burghardt, 2004; Studie 3). Daher ist anzunehmen, dass der vorliegende Befund auf Schwierigkeiten im inhaltlichen Verständnis des ARJS-STV-G-Merkmals Komplikationen zurückzuführen ist. Diese Vermutung wird dadurch gestützt, dass sich für die Beurteilung dieses Aussagemerkmals keine ausreichende Inter-Rater-Reliabilität feststellen ließ.

Für die Beurteilung der Interviews wiederum ergab sich eine signifikante Interaktion zwischen dem Wahrheitsstatus und der Vorbereitungszeit hinsichtlich des ARJS-STV-G-Merkmals persönliche Bedeutsamkeit. Es wurden nur dann erwartungsgemäße Unterschiede zwischen wahren und erfundenen Aussagen

vorgefunden, wenn sie unvorbereitet erfolgten. Für vorbereitete Aussagen wiederum waren keine Unterschiede in Abhängigkeit vom Wahrheitsstatus festzustellen. Dies war darauf zurückzuführen, dass bei erfundenen Aussagen die Gelegenheit zur Vorbereitung genutzt wurde, um die persönliche Bedeutsamkeit des Ereignisses herauszustellen.

Im Rahmen der vorliegenden Untersuchung wurden die Probanden instruiert, persönlich bedeutsame Lebensereignisse zu schildern. Personen, die Erfundenes schilderten, nutzten diese Instruktion möglicherweise als Hinweis, wie sie ihre Aussage überzeugend gestalten können. Wurde ihnen Gelegenheit zur Vorbereitung gewährt, setzten sie diesen Hinweis offenbar erfolgreich um.

Auch Sporer (1998) fand für die Skala persönliche Signifikanz der ARJS-Langform eine höhere Validität bei vorbereiteten Aussagen im Vergleich zu unvorbereiteten. Für das in der vorliegenden Untersuchung verwendete Stimulusmaterial sind jedoch widersprüchliche Befunde bei den Analysen anhand der ARJS-STV-G und ARJS-Langform festzustellen. Bei den Beurteilungen anhand der Langform wurde die Validität der Skala persönliche Signifikanz nicht durch die Gelegenheit zu Vorbereitung beeinflusst (vgl. Studie 3; Sporer & Burghard, 2004). Möglicherweise ist die Langform gegenüber potenziell validitätsreduzierenden Einflüssen der Gelegenheit zur Vorbereitung robuster als die Kurzform.

Insgesamt bleibt festzuhalten, dass hypothesengemäß Einflüsse der Vorbereitungszeit auf die ARSJ-STV-G-Beurteilungen erkennbar waren. Allerdings ließ sich weder der Haupteffekt der Vorbereitungszeit noch die Interaktion mit dem Wahrheitsstatus multivariat absichern. Daher könnten die ARJS-STV-G-Merkmale unabhängig von der Gelegenheit zur Vorbereitung einer Aussage valide sein.

Effekte der Aussageform

Die Interviews wiesen eine höhere Aussagequalität hinsichtlich einzelner ARJS-STV-G-Merkmale auf als die freien Berichte. Die stärksten Effekte ergaben sich für die Merkmale zeitliche Details, räumliche Details, Erinnerungslücken zugeben und Details. Dies ist vermutlich darauf zurückzuführen, dass die

Interviewerin explizit nach spezifischen Details fragte und oftmals auch nach konkreten Namen, die leicht in Vergessenheit geraten können. Des Weiteren zeigten sich Effekte geringer Größenordnung hinsichtlich der Merkmale Klarheit und Lebendigkeit, Realismus und logische Stuktur sowie überflüssige Details. Auch dies könnte durch konkrete Nachfragen der Interviewerin zu begründen sein. So wurden die Probanden zuweilen dazu aufgefordert, sich die Situationen bildlich vorzustellen und zuvor Erwähntes weiter auszuführen.

In der aussagepsychologischen Literatur wird empfohlen, freie Berichte für die inhaltliche Glaubhaftigkeitsanalyse zu verwenden (z.B. Steller & Köhnken, 1989). In der vorliegenden Untersuchung war jedoch keine Wechselwirkung zwischen dem Wahrheitsstatus und der Aussageform festzustellen. Entsprechend ist anzunehmen, dass es auch auf spezifische Nachfragen hin besser möglich ist, sich an einzelne Aspekte wahrheitsgemäß zu erinnern als diese frei zu erfinden. Zumindest bildeten die ARJS-STV-G unabhängig von der Aussageform erwartungsgemäße Mittelwertsunterschiede zwischen wahren und erfundenen Aussagen ab. Infolgedessen scheint deren Validität auf Interviews generalisierbar zu sein. Problematisch wäre es allerdings, spezifische Frragen nach einzelnen ARJS-Merkmalen zu stellen (Sporer & Walther, 2006).

Effekte der Valenz der geschilderten Ereignisse

Schließlich wurde angenommen, dass die ARJS-STV-G-Merkmale bei Aussagen zu negativen im Vergleich zu positiven Ereignissen deutlicher ausgeprägt sind. Für die freien Berichte ließ sich kein Effekt der Valenz des geschilderten Ereignisses auf die Aussagequalität nachweisen. Interviews zu negativen Ereignissen zeigten jedoch erwartungsgemäß eine höhere Aussagequalität als zu positiven Ereignissen. Allerdings ist darauf hinzuweisen, dass für die Analysen Ereignisse als positiv bezeichnet wurden, die anhand der verwendeten Beurteilungsskalen eher als neutral eingestuft wurden. Infolgedessen überrascht es kaum, dass bei verhältnismäßig positiven Ereignissen seltener Emotionen und Gefühle geschildert wurden als bei

negativen. Weitere Mittelwertsunterschiede zeigten sich vor allem für gedächtnistheoretisch fundierte Glaubhaftigkeitsmerkmale. So ergaben sich mittlere bis geringe Effekte hinsichtlich der ARJS-STV-G-Merkmale Gedanken, Gedächtnisprozesse und Memorieren, Klarheit und Lebendigkeit sowie persönliche Bedeutsamkeit. Auch das aus den forensischen Kriterien abgeleitete Merkmal nonverbale und verbale Interaktionen war häufiger bei negativen als bei positiven Ereignissen vorzufinden. Hingegen variierten Merkmale, deren Validität sozialpsychologisch begründet wird, nicht in Abhängigkeit von der Valenz der geschilderten Ereignisse. Dies verweist darauf, dass Personen unabhängig von der Valenz des geschilderten Ereignisses um eine überzeugende Darstellung bemüht sind.

Im Gegensatz dazu wiesen Selbstbeurteilungsstudien kaum Qualitätsunterschiede zwischen negativen und neutralen Erinnerungen nach (D'Argembeau, Comblain & van der Linden, 2003; Schaefer & Philippot, 2005). Möglicherweise hat die kognitive Anstrengung, die Probanden im Rahmen eines Experiments aufwenden, einen entscheidenden Einfluss auf die Erinnerungs- und Aussagequalität. Nach Conway und Pleydell-Pearce (2000) sind Personen darum bemüht, negative Ereignisse zu vergessen. Dies könnte eine geringere Erinnerungsqualität negativer Ereignisse im Vergleich zu neutralen begründen. Wird die mit negativen Ereignissen verbundene Angst jedoch durch kognitive Anstrengung überwunden, könnte sich infolgedessen auch die Aussagequalität erhöhen (vgl. Barnier, Sharman, McKay & Sporer, 2005). Bei der Darstellung neutraler Ereignisse sind hingegen keine entsprechenden Bemühungen notwendig. Dies könnte eine höhere Aussagequalität für negative als für neutrale Ereignisse erklären.

Für die vorliegende Untersuchung ist durchaus von einer Bereitschaft der Probanden auszugehen, die mit den Ereignissen verbundenen negativen Gefühle zu überwinden. Zum einen wurde die Valenz der zu schildernden Ereignisse nicht experimentell manipuliert. Vielmehr wurde den Probanden freigestellt, zu welchem

Ereignis sie sich äußern möchten. Die Entscheidung für ein Ereignis ging vermutlich auch mit der Bereitschaft einher, sich detailliert daran zu erinnern. Zum anderen sollten sich die Probanden die entsprechenden Ereignisse nicht nur vorstellen (D'Argembeau et al., 2003), sondern sich ausführlich mündlich dazu äußern und konkrete Nachfragen im Rahmen von Interviews beantworten.

Möglicherweise trugen diese beiden Aspekte des Untersuchungsdesigns dazu bei, dass die Probanden kognitive Anstrengungen aufwendeten, um die mit negativen Ereignissen verbundene Angst zu überwinden. Infolgedessen konnten die Fremdbeurteilungen anhand der ARJS-STV-G erwartungsgemäße Qualitätsunterschiede zwischen negativen und verhältnismäßig positiven Ereignissen abbilden (vgl. auch Studie 3). Zudem wird der vorliegende Befund durch eine Untersuchung von Barnier et al. (2005) gestützt, die ebenfalls höhere Fremdbeurteilungen der Aussagequalität für negative Ereignisse im Vergleich zu positiven berichteten.

Des Weiteren wurden Wechselwirkungseffekte zwischen dem Wahrheitsstatus und der Valenz des geschilderten Ereignisses untersucht. Dadurch sollte geklärt werden, ob die ARJS-STV-G auch bei positiven bzw. neutralen Ereignissen Qualitätsunterschiede zwischen wahren und erfundenen Aussagen abbilden. Entsprechende Interaktionseffekte zwischen dem Wahrheitsstatus und der Valenz der geschilderten Ereignisse waren lediglich für die Interviews festzustellen. Für zeitliche Details ergaben sich nur bei negativen Ereignissen erwartungsgemäße Mittelwertsunterschiede zwischen wahren und erfundenen Aussagen. Im Gegensatz dazu erlaubten es die beiden Merkmale Gedanken und räumliche Details nur bei positiven Ereignissen, wahre und erfundene Aussagen zu unterscheiden. Infolgedessen erscheint die Schlussfolgerung nicht haltbar, dass die Negativität des Ereignisses Voraussetzung für die Validität inhaltlicher Aussagemerkmale sei (Landry & Brigham, 1992; Steller & Köhnken, 1989). Die differentielle Validität einzelner

Merkmale ergibt sich vermutlich weniger aus der allgemeinen Valenz des geschilderten Ereignisses, sondern vielmehr aus der spezifischen Thematik.

Weiterführender Forschungsbedarf und praktische Implikationen

Die vorliegende Untersuchung erzielte vielversprechende Ergebnisse für eine ökonomische Kurzform der ARJS. Raterinnen ohne Vorwissen in der Glaubhaftigkeitsanalyse wendeten die ARJS-STV-G nach einer kurzen Informationsveranstaltung reliabel an. Allerdings konnte eine gute Inter-Rater-Reliabilität für die meisten Merkmale lediglich aufgrund der Zusammenfassung mehrerer unabhängiger Beurteilungen gesichert werden. Folglich ist davon abzuraten, nur einen Rater für die Aussageanalyse einzusetzen. Vielmehr sollten verschiedene Personen über die ARJS-STV-G informiert werden und ihre Beurteilungen zusammengetragen werden.

Zudem erzielten wahre Aussagen erwartungsgemäß höhere ARJS-STV-G-Beurteilungen als erfundene. Die Validität der Glaubhaftigkeitsmerkmale wurde allerdings deutlich reduziert, wenn die Anzahl der Wörter als Kovariate berücksichtigt wurde. Für Rater ohne fundiertes Hintergrundwissen in der Glaubhaftigkeitsdiagnostik erscheint es demnach schwierig, die Qualität einer Aussage unabhängig von deren Länge einzuschätzen. Andererseits kann jedoch die Anzahl der Wörter ebenfalls ein Indikator für den Wahrheitsstatus einer Aussage sein (z.B. Strömwall, Bengtsson, Leander & Granhag, 2004; Vrij, Edward, Roberts & Bull, 2000). Das eigentliche Ziel der Glaubhaftigkeitsanalyse, wahre und erfundene Aussagen voneinander zu unterscheiden, wäre folglich nicht zwangsläufig gefährdet. Allerdings bleibt darauf hinzuweisen, dass metaanalytische Untersuchungen nur tendenziell kürzere erfundene als wahre Aussagen fanden (z.B. DePaulo et al., 2003; Sporer & Schwandt, 2006). Dieser Befund scheint jedoch von der Operationalisierung der Aussagelänge abhängig zu sein und zudem bleibt seine praktische Relevanz aufgrund der geringen Effektstärken fraglich.

Weder die Vorbereitungszeit noch die Aussageform interagierten mit dem Wahrheitsstatus. Die ARJS-STV-G waren also unabhängig von diesen beiden Faktoren valide. Hingegen moderierte die Valenz der geschilderten Ereignisse die Validität der ARJS-STV-G-Merkmale. Dabei fiel die Validität zeitlicher Details bei negativen Ereignissen besser aus, die Validität räumlicher Details und Gedanken jedoch bei positiven Ereignissen. Demnach scheint die Anwendung der ARJS-STV-G nicht auf negative Ereignisse beschränkt zu sein (vgl. Steller, 1989). Dennoch erscheint es sinnvoll, die differentielle Validität einzelner ARJS-STV-G-Merkmale in weiteren Studien zu überprüfen. Dabei könnten neben der Valenz auch andere Merkmale der geschilderten Ereignisse bedeutsam sein. Beispielsweise wäre es denkbar, dass die Zeit, die seit dem Ereignis vergangen ist, die Validität der ARJS-STV-G-Urteile moderiert (vgl. Sporer & Sharman, 2006).

Insgesamt verweisen die vorliegenden Befunde darauf, dass es nützlich sein könnte, neben psychologischen Sachverständigen auch andere Berufsgruppen über inhaltliche Glaubhaftigkeitsmerkmale zu informieren. Potenziell wäre eine Anwendung in sämtlichen Bereichen denkbar, in denen die Glaubhaftigkeit von Aussagen darüber entscheidet, ob die Hypothese einer Falschaussage weiter überprüft wird. Solche vorläufigen Entscheidungen werden bislang oftmals intuitiv getroffen. Dies gilt beispielsweise für die Ermittlungstätigkeiten von Polizeibeamten und für die Bearbeitung von Schadensansprüchen durch Sachbearbeiter in Versicherungsunternehmen.

Wenn jedoch Richter oder Staatsanwälte für gerichtliche Verhandlungen eine Glaubhaftigkeitsanalyse anfordern, sollte diese den Experten vorbehalten bleiben. Schließlich bedeutet die Glaubhaftigkeitsdiagnostik im Einzelfall weitaus mehr, als die Qualität einer spezifischen Aussage einzuschätzen. Vielmehr sind zusätzlich fundierte Diagnosen der Aussagetüchtigkeit und –entwicklung vorzunehmen. Zudem zeigen die vorliegenden Untersuchungsbefunde lediglich, dass wahre Aussagen höhere ARJS-STV-G-Beurteilungen erzielen als erfundene. Daraus lässt sich schlussfolgern, dass die inhaltlichen Aussagemerkmale valide

sind. Hingegen bleibt unklar, ob auf der Grundlage dieser Beurteilungen auch valide Entscheidungen hinsichtlich der Glaubhaftigkeit von Aussagen getroffen werden. Im Rahmen der nachfolgenden Studie wurde daher die Güte der endgültigen Glaubhaftigkeitsurteile überprüft.

Literatur

Alonso-Quecuty, M. (1992). Deception detection and reality monitoring: A new anser to an old question? In F. Loesel, D. Bender, & T. Bliesener (Eds.), Psychology and Law: International perspectives (pp. 228-332). Berlin: Walter de Gruyter.

Barnier, A. J., Sharman, S. J., McKay, L., & Sporer, S. L. (2005) Discriminating adults' genuine, imagined, and deceptive accounts of positive and negative childhood events. Applied Cognitive Psychology, 19, 985-1001

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. Psychological Review, 107, 261-288.

Cramer, M. (2005). Can we train people to detect deception? Unpublished master's thesis, Justus-Liebig-Universität Gießen, Germany.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. Psychological Bulletin, 129, 74-112.

D'Argembeau, A., Comblain, C., & van der Linden, M. (2003). Phenomenal characteristics of autobiographical memories for positive, negative and neutral events. Applied Cognitive Psychology, 17, 281-294.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. Psychological Review, 88, 67-85.

Landry, K., & Brigham, J. C. (1992). The effect of training in criteria-based content analysis on the ability to detect deception in adults. Law and Human Behavior, 16, 663-675.

Schaefer, A., & Philippot, P. (2005). Selective effects of emotion on the phenomenal characteristics of autobiographical memories. Memory, 13, 148-160.

Sporer, S. L. (1996/1998/2004). The Aberdeen Report Judgment Scales (ARJS). Definitions and answer sheets. Unpublished Questionnaire. University of Aberdeen, Scotland; University of Giessen, Germany.

Sporer, S. L. (1997). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experiences events. Applied Cognitive Psychology, 11, 373-397.

Sporer, S. L. (1998, March). Detecting deception with the Aberdeen Report Judgment Scales (ARJS): Theoretical development reliability and validity. Paper presented at the Biennial Meeting of the American Psychology-Law Society in Redondo Beach, CA.

Sporer, S. L., & Burghardt, S. E. (2004, March). Truth detection with the Aberdeen Report Judgment Scales: The role of planning and rehearsal. Paper presented at the Biennial Meeting of the American Psychology-Law Society in Phoenix, AZ.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. Applied Cognitive Psychology, 20, 421-446.

Sporer, S. L., & Sharman, S. J. (2006). Should I believe this? Reality monitoring of accounts of self-experienced and invented recent and distant autobiographical events. Applied Cognitive Psychology, 20, 1-18.

Sporer, S. L., & Masip, J. (2007). Guidance to detect deception by content cues: Self-efficacy of liars with different types of lies. Final Report to the DAAD. Giessen, Germany, & Salamanca, Spain.

Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille (Ed.), Credibility Assessment (pp. 135-154). Deventer: Kluwer Academic Publishers.

Steller, M., & Köhnken, G. (1989). Criteria-based statement analysis. In D. C. Raskin (Ed). Psychological methods in criminal investigation and evidence (pp. 217-245). New York: Springer.

Strömwall, L. A., Bengtsson, L., Leander, L., & Granhag, P. A. (2004). Assessing children's statements: The impact of a repeated experience on CBCA and RM ratings. Applied Cognitive Psychology, 18, 653-668.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. Journal of Nonverbal Behavior, 24, 239-263.

Anhang 4a

Inter-Rater-Reliabilitäten (Pearson-Korrelationen) der 17 ARJS-STV-G-Merkmale für die freien Berichte (B01.01 – B11.16)

Merkmale	r_{AB}	r_{CD}	r_{EF}	r_{GH}	M(r)	$r_{SP(2)}$
Negative Äußerungen über sich selbst	.43	.55	.30	.33	.41	.58
Spontane Korrekturen	.07	.44	-.07	-.07	.10	.18
Erinnerungslücken zugeben	.64	.56	.51	.27	.51	.67
Komplikationen	.10	.07	.38	.45	.26	.41
Ungewöhnliche Details	.38	.23	.00	.10	.18	.31
Überflüssige Details	.54	.14	.14	.22	.27	.43
Nonverbale und verbale Interaktionen	.59	.87	.64	.67	.71	.83
Emotionen und Gefühle	.53	.32	.38	.59	.46	.63
Sensorische Eindrücke	.58	.70	-	.45	.58	.74
Gedächtnisprozesse und Memorieren	.34	.00	.24	.14	.18	.31
Gedanken	.71	.26	.20	.53	.45	.62
Räumliche Details	.29	.36	.36	.48	.37	.54
Zeitliche Details	.62	.35	.26	.59	.47	.64
Details	.46	.36	.12	.52	.37	.54
Persönliche Bedeutsamkeit	.42	.41	-.02	.60	.37	.54
Klarheit und Lebendigkeit	.47	.18	.19	.22	.27	.43
Realismus und logische Struktur	.50	.00	.14	.26	.23	.38

Anm. N = 176. Alle Raterinnen beurteilten jeweils ein Viertel der freien Berichte. - = für das Raterinnenpaar E—F ließ sich keine korrelative Übereinstimmung für sensorische Eindrücke berechnen, da Raterin E alle freien Berichte mit 1 bewertete.

Anhang 4b

Inter-Rater-Reliabilitäten (Pearson-Korrelationen) der 17 ARJS-STV-G-Merkmale für die Interviews (C01.01 – C11.16)

	Raterinnen A - D								Raterinnen E- H							
Items	r_{AB}	r_{AC}	r_{AD}	r_{BC}	r_{BD}	r_{CD}	M(r)	$r_{SP(4)}$	r_{EF}	r_{EG}	r_{EH}	r_{FG}	r_{FH}	r_{GH}	M(r)	$r_{SP(4)}$
01	.37	.52	.35	.48	.48	.44	.44	.76	.12	.20	.26	-.03	.46	.05	.18	.47
02	.25	.37	.13	.33	.26	.33	.28	.61	-.07	.09	-.01	.40	.37	.16	.16	.43
03	.57	.65	.60	.62	.62	.49	.59	.85	.59	.25	.54	.25	.63	.32	.45	.76
04	-.14	.17	.14	.14	.24	.30	.14	.40	.33	.13	.47	.34	.31	.39	.33	.66
05	.38	.26	.30	.19	.36	.15	.27	.60	.26	.10	.12	.11	.31	.14	.17	.46
06	.42	.33	.36	.35	.32	.50	.38	.71	.11	.31	.40	.36	.21	.25	.27	.60
07	.63	.57	.56	.62	.74	.71	.64	.88	.61	.73	.56	.60	.65	.50	.61	.86
08	.52	.52	.43	.44	.39	.40	.45	.77	.43	.56	.50	.45	.57	.67	.53	.82
09	.47	.34	.43	.44	.58	.55	.47	.78	.61	.55	.34	.35	.49	.11	.42	.74
10	.19	.16	.26	.16	.39	.17	.22	.54	.29	-.06	.05	.29	.27	.04	.15	.41
11	.71	.42	.42	.46	.50	.37	.49	.79	.43	.49	.31	.41	.25	.41	.39	.72
12	.45	.45	.31	.65	.27	.28	.41	.74	.30	.42	.27	.45	.32	.50	.38	.71
13	.58	.57	.13	.61	.37	.22	.43	.75	.35	.43	.41	.42	.50	.50	.44	.76
14	.44	.38	.43	.47	.55	.56	.47	.78	.22	.10	.33	.42	.55	.54	.37	.70
15	.43	.56	.53	.42	.56	.51	.50	.80	.23	.21	.45	.38	.25	.54	.35	.68
16	.31	.24	.48	.24	.54	.42	.38	.71	-.18	.26	.30	-.45	.13	.12	.02	.09
17	.08	.13	.19	.14	.39	.22	.19	.49	.16	.12	.35	.04	.23	.08	.17	.45

Anm. N = 176. Alle Raterinnen beurteilten jeweils die Hälfte der Interviews anhand der ARJS-STV-G. Items: 01 Negative Äußerungen über sich selbst, 02 Spontane Korrekturen, 03 Erinnerungslücken zugeben, 04 Komplikationen, 05 Ungewöhnliche Details, 06 Überflüssige Details, 07 Nonverbale und verbale Interaktionen, 08 Emotionen und Gefühle, 09 Sensorische Eindrücke, 10 Gedächtnisprozesse und Memorieren, 11 Gedanken, 12 Räumliche Details, 13 Zeitliche Details, 14 Details, 15 Persönliche Bedeutsamkeit, 16 Klarheit und Lebendigkeit, 17 Realismus und logische Struktur. M(r) basiert auf den unabhängigen Beurteilungen durch vier Raterinnen. $r_{SP(4)}$ entspricht Spearman-Brown-korrigierten Werten für vier Rater.

Anhang 4c

Mittelwertsunterschiede zwischen wahren und erfundenen Aussagen hinsichtlich der 17 ARJS-STV-G-Merkmale unter Kontrolle der Anzahl der Wörter für die freien Berichte (B01.01-B11.16)

Merkmale	$M_{erfunden}$	M_{wahr}	F(1,171)	p	r
Negative Äußerungen über sich selbst	1.64	1.99	4.11	**.044**	**.15**
Spontane Korrekturen	1.50	1.46	0.06	.810	-.02
Erinnerungslücken zugeben	1.89	1.87	0.01	.933	-.01
Komplikationen	2.10	2.25	0.57	.452	.06
Ungewöhnliche Details	1.92	2.28	3.79	.053	.15
Überflüssige Details	2.68	3.06	2.91	.090	.13
Nonverbale und verbale Interaktionen	3.94	4.14	0.62	.433	.06
Emotionen und Gefühle	4.69	5.00	2.78	.097	.13
Sensorische Eindrücke	1.85	1.92	0.09	.759	.02
Gedächtnisprozesse und Memorieren	3.52	3.64	0.30	.584	.04
Gedanken	4.35	4.66	1.66	.199	.10
Räumliche Details	3.35	3.18	0.95	.330	-.07
Zeitliche Details	3.90	4.22	1.99	.160	.11
Details	3.88	4.21	5.41	**.021**	**.18**
Persönliche Bedeutsamkeit	5.16	5.05	0.29	.593	-.04
Klarheit und Lebendigkeit	5.27	5.58	4.80	**.030**	**.17**
Realismus und logische Struktur	5.60	5.87	3.85	.051	.15

Anm. N = 176; 7-stufige Ratingskalen. Positive Effektstärken r signieren eine höhere Ausprägung in wahren Berichten.

Anhang 4d

Mittelwertsunterschiede zwischen wahren und erfundenen Aussagen hinsichtlich der 17 ARJS-STV-G Merkmale unter Kontrolle der Anzahl der Wörter für die Interviews (C01.01-C11.16)

Merkmale	$M_{erfunden}$	M_{wahr}	F(1,171)	p	r
Negative Äußerungen über sich selbst	1.73	1.93	1.65	.201	.10
Spontane Korrekturen	1.51	1.74	2.73	.101	.13
Erinnerungslücken zugeben	3.11	3.12	0.00	.987	.00
Komplikationen	2.02	2.31	3.20	.076	.14
Ungewöhnliche Details	1.87	2.30	8.74	**.004**	**.22**
Überflüssige Details	2.96	3.42	13.53	**.000**	**.27**
Nonverbale und verbale Interaktionen	3.69	4.33	9.17	**.003**	**.23**
Emotionen und Gefühle	4.71	4.96	2.20	.140	.11
Sensorische Eindrücke	2.12	1.95	0.82	.367	-.07
Gedächtnisprozesse und Memorieren	3.70	3.81	0.63	.429	.06
Gedanken	4.50	4.72	1.93	.167	.11
Räumliche Details	4.11	4.24	0.54	.462	.06
Zeitliche Details	5.11	5.29	1.14	.287	.08
Details	4.27	4.57	10.53	**.001**	**.24**
Persönliche Bedeutsamkeit	5.10	5.08	0.02	.895	-.01
Klarheit und Lebendigkeit	5.49	5.77	8.49	**.004**	**.22**
Realismus und logische Struktur	5.72	6.09	16.40	**.000**	**.30**

Anm. N = 176; 7-stufige Ratingskalen. Positive Effektstärken r signieren eine höhere Ausprägung in wahren Interviews.

STUDIE 5

Anleitung anhand der Aberdeen Report Judgment Scales--Short Training Version--German: Effekte auf die Urteilsgüte, -neigung und –sicherheit

Die beiden vorangegangenen Studien haben gezeigt, dass die ARJS-Merkmale zu einer validen Unterscheidung von wahren und erfundenen Aussagen beitragen können. Dies gilt sowohl für die Langform als auch für die ökonomischere Kurzform. Allerdings bleibt bislang ungeklärt, ob Beurteiler dieses Wissen über empirisch fundierte Aussagemerkmale auch nutzen, um richtige Entscheidungen hinsichtlich der Glaubhaftigkeit von Aussagen zu treffen. Diese Frage nach der Richtigkeit des subjektiven Glaubhaftigkeitsurteils bzw. nach der Urteilsgüte ist Gegenstand der vorliegenden Untersuchung.

Quantifizierung der Urteilsgüte

Um die Urteilsgüte im Labor zu ermitteln, werden Personen mehrere Aussagen verschiedener Stimuluspersonen zur Beurteilung vorgelegt. In der Regel wird manipuliert, ob diese Aussagen wahr oder erfunden sind. Der tatsächliche Wahrheitsstatus der zu beurteilenden Basissachverhalte stellt demnach eine dichotome Variable dar. Die Beurteiler geben entweder dichotome Glaubhaftigkeitsurteile ab, oder schätzen den Wahrheitsgehalt jeder Aussage auf mehrstufigen Skalen ein. Es lassen sich verschiedene Indizes zur Bestimmung der Urteilsgüte ableiten. Im Folgenden werden zunächst für die Auswertung relevante Begrifflichkeiten definiert und erläutert. Anschließend wird dargestellt, wie prozentuale Urteilsrichtigkeiten, signalentdeckungstheoretische Indizes und standardisierte Mittelwertsdifferenzen berechnet werden und zu interpretieren sind.

Entscheidungsausgänge und relevante Einflussfaktoren

Dichotome Glaubhaftigkeitsurteile lassen sich direkt mit dem tatsächlichen Wahrheitsstatus abgleichen. Die sich daraus ergebenden vier möglichen Entscheidungsausgänge sind in Tabelle 5.1 dargestellt.

Tabelle 5.1
Mögliche Entscheidungsausgänge bei der Beurteilung wahrer und erfundener Aussagen.

Tatsächlicher Wahrheitsstatus	Subjektives Glaubhaftigkeitsurteil		Summe
	Glaubhaft (G)	Nicht-Glaubhaft (NG)	
Wahr (W)	Treffer (Hits, H)	Falsche Zurückweisungen (Misses, M)	n(W) = H+M
Nicht-wahr (NW) bzw. erfunden	Falsche Alarme (False Alarms, F)	Korrekte Zurückweisungen (Correct Rejections, CR)	n(NW) = CR + F
Summe	n(G) = H+F	n(NG)= CR+M	N = H+M+CR+F

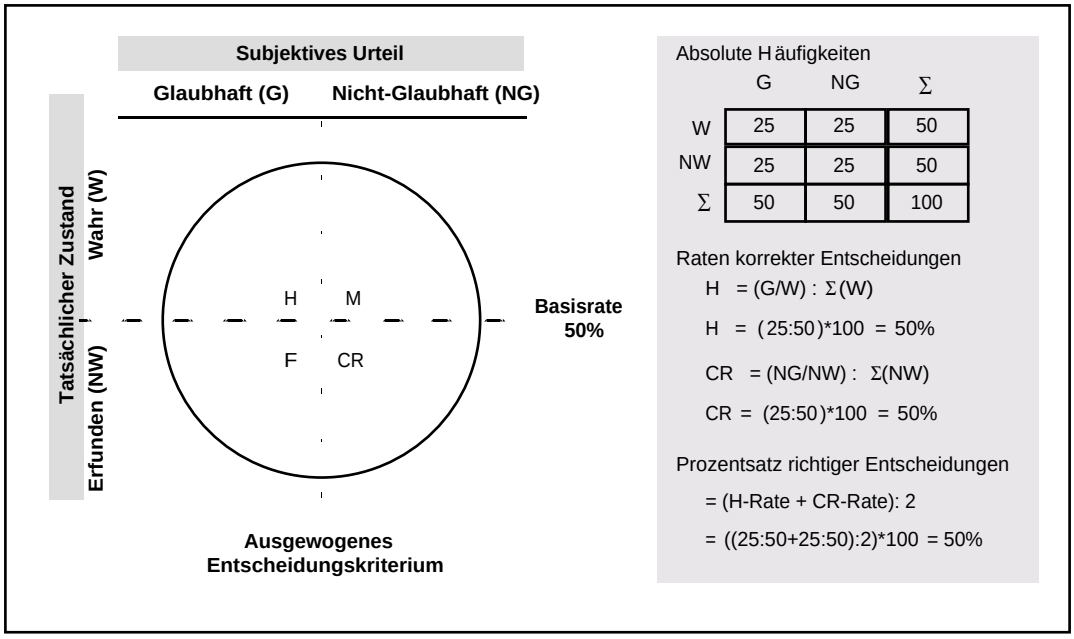
In der Täuschungsliteratur besteht keineswegs Einigkeit darüber, wie diese zu bezeichnen sind. Für die inhaltliche Aussageanalyse werden überwiegend Merkmale verwendet, deren Vorhandensein die Hypothese eines persönlichen Erlebnisbezugs stärkt. Daher wird in der vorliegenden Arbeit von Treffern gesprochen, wenn tatsächlich wahre Aussage als glaubhaft eingeschätzt wurden.[1] Korrekte Zurückweisungen beziehen sich entsprechend auf tatsächlich erfundene Aussagen, die als nicht-glaubhaft beurteilt wurden. Falsche Entscheidungen liegen vor, wenn entweder tatsächlich erfundene Aussagen als glaubhaft beurteilt wurden (falsche Alarme) oder tatsächlich wahre als nicht-glaubhaft (falsche Zurückweisungen).

Die Abbildungen[2] und fiktiven Rechenbeispiele 5.1a bis 5.1e veranschaulichen die Zusammenhänge zwischen Treffern, korrekten Zurückweisungen, falschen Alarmen und falschen Zurückweisungen. Zudem wird der Einfluss der Beurteilungsfähigkeit, der Basisrate und der Urteilsneigung auf die vier Entscheidungsausgänge illustriert. Dabei bezieht sich die Basisrate auf den Anteil tatsächlich wahrer an allen zu beurteilenden Aussagen. Die Urteilsneigung repräsentiert, wieviele Aussagen als glaubhaft bzw. als nicht-glaubhaft beurteilt wurden.

---

[1] Diese Kategorisierung erlaubt es, die vorliegenden Untersuchungsbefunde direkt mit den Forschungsarbeiten von Bond und DePaulo (2006) sowie von Sporer (2004) zu vergleichen. Im Gegensatz dazu definierten andere Autoren Treffer als erfundene Aussagen, die richtig beurteilt wurden (z.B. Bond & Lee, 2005; G. D. Bond, Malloy et al., 2005; G. D. Bond, Thompson et al., 2005; Meissner & Kassin, 2002; Vrij, Mann, Kristen & Fisher, 2007). Ihre Befunde wurden umgerechnet, so dass sie mit der in der vorliegenden Arbeit verwendeten Definition im Einklang stehen.

[2] Eine ähnliche Darstellungsweise findet sich beispielsweise bei Schuler (2004, S. 331). Sie wurde lediglich auf das Entdecken von Täuschung übertragen.



Absolute Häufigkeiten

	G	NG	Σ
W	25	25	50
NW	25	25	50
Σ	50	50	100

Raten korrekter Entscheidungen

H = (G/W) : Σ(W)

H = (25:50)*100 = 50%

CR = (NG/NW) : Σ(NW)

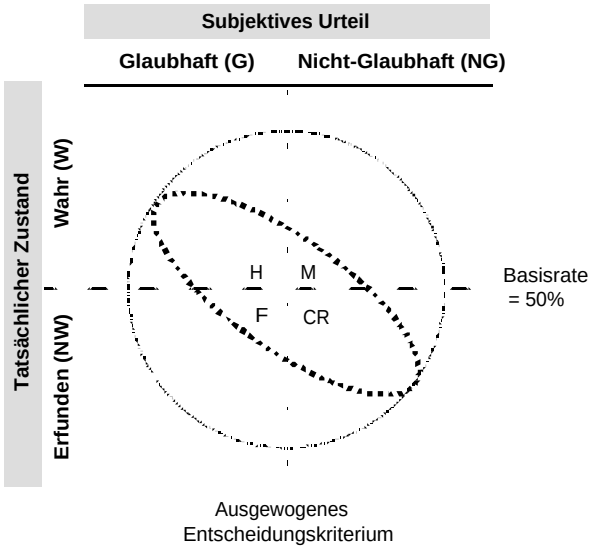
CR = (25:50)*100 = 50%

Prozentsatz richtiger Entscheidungen

= (H-Rate + CR-Rate): 2

= ((25:50+25:50):2)*100 = 50%

Abbildung und Rechenbeispiel 5.1a: Ableitung des Prozentsatzes richtiger Entscheidungen als Indikator der Urteilsgüte.



Absolute Häufigkeiten

	G	NG	Σ
W	40	10	50
NW	10	40	50
Σ	50	50	100

Raten korrekter Entscheidungen

H = (25:50)*100 = 50%

CR = (25:50)*100 = 50%

Prozentsatz richtiger Entscheidungen

= ((40:50+40:50):2)*100 = 80%

Abbildung und Rechenbeispiel 5.1b: Urteilsgüte bei überzufälliger Beurteilungsfähigkeit, einer Basisrate von 50% und einem ausgewogenen Entscheidungskriterium.

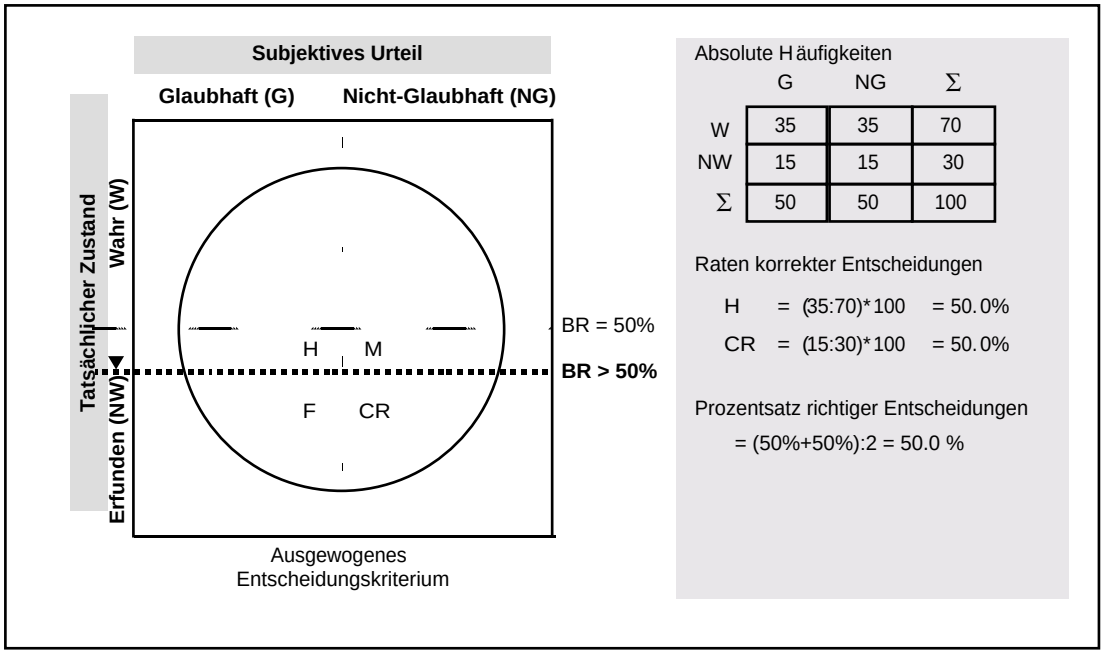


Abbildung und Rechenbeispiel 5.1c: Urteilsgüte bei einer Basisrate (BR) von 70%, einer zufälligen Beurteilung und einem ausgewogenen Entscheidungskriterium.

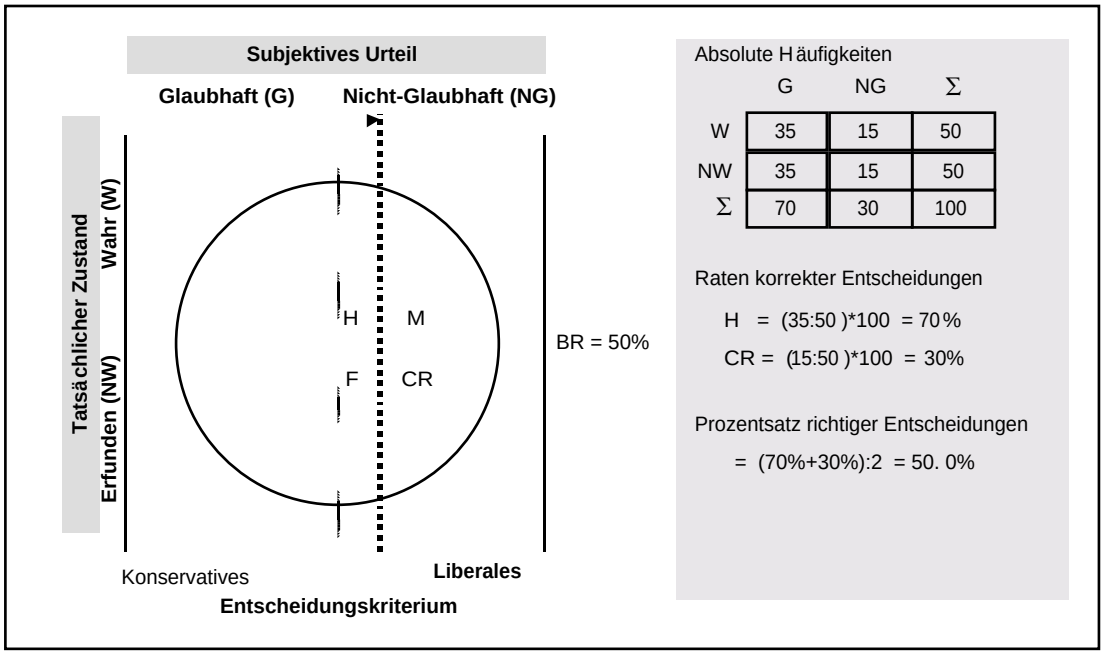


Abbildung und Rechenbeispiel 5.1d: Urteilsgüte bei einem liberalen Entscheidungskriterium, einer zufälligen Beurteilung und einer Basisrate (BR) von 50%.

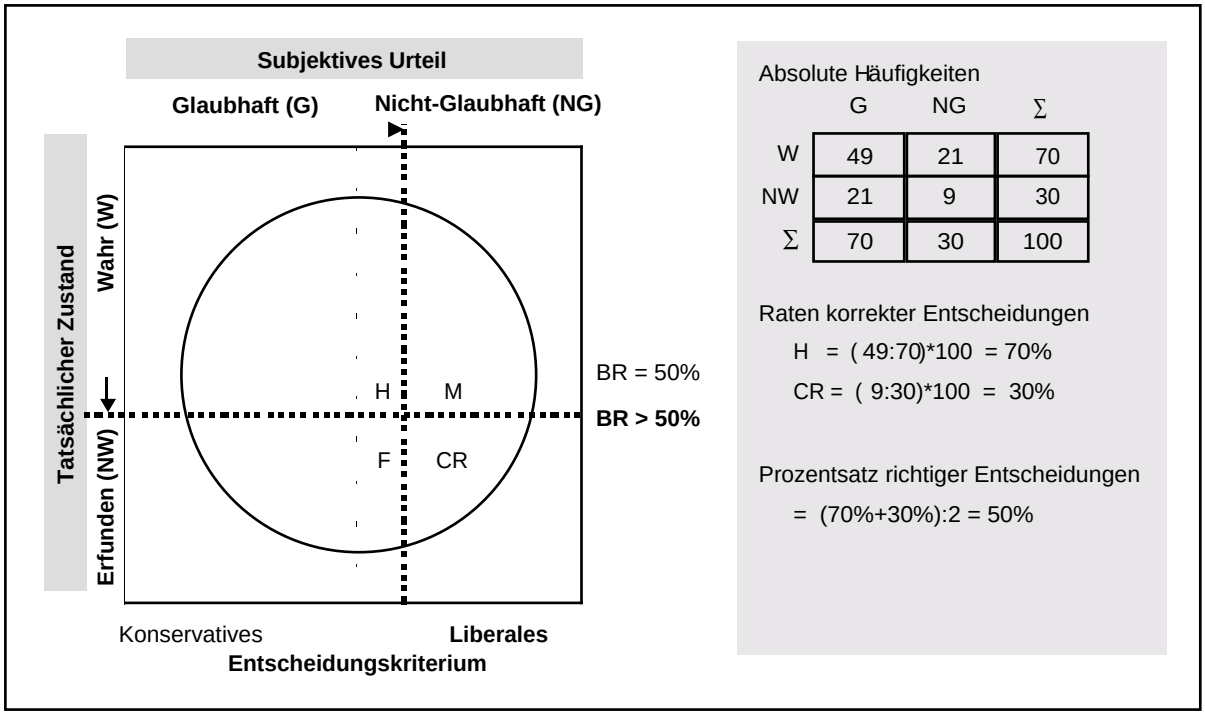


Abbildung und Rechenbeispiel 5.1e: Urteilsrichtigkeit für eine Basisrate (BR) von 70%, einem liberalen Entscheidungskriterium und einer zufälligen Beurteilung.

Die Basisrate und Urteilsneigung sind in den Abbildungen und Rechenbeispielen 5.1a und 5.1b gleich gesetzt. Hier wurde davon ausgegangen, dass gleich viele wahre und erfundene Aussagen bewertet wurden. Wie es in Laborstudien üblich ist, liegt die Basisrate demnach bei 50%. Des Weiteren wurde vorausgesetzt, dass sich die Beurteiler gleichermaßen häufig für bzw. gegen die Glaubhaftigkeit entscheiden. Sie verwenden damit ein ausgewogenes Entscheidungskriterium. Die beiden Beispiele unterscheiden sich jedoch hinsichtlich der Beurteilungsfähigkeit. So zeigen Abbildung und Rechenbeispiel 5.1a, dass rein zufällige Urteile in einer Entscheidungsrichtigkeit von 50% resultieren. Hingegen illustrieren Abbildung und Rechenbeispiel 5.1b eine gute Beurteilungsfähigkeit. Diese spiegelt sich in einer (durch die Ellipse symbolisierten) bedeutsamen Korrelation zwischen dem objektiven Wahrheitsstatus und dem subjektiven Glaubhaftigkeitsurteil wider. Es resultiert eine bessere Urteilsrichtigkeit als bei einer zufälligen Beurteilung. Dies ist darauf

zurückzuführen, dass die richtigen Entscheidungen zunehmen, während die beiden Fehlerarten abnehmen.

Abbildung und Rechenbeispiel 5.1c zeigen die Auswirkungen einer Basisratenerhöhung auf die Verteilung der vier Entscheidungsausgänge. Dabei wurde davon ausgegangen, dass Beurteiler keine überzufällige Fähigkeit aufweisen, wahre und erfundene Aussagen zu unterscheiden. Sie beurteilen diese jedoch gleich häufig als glaubhaft und nicht-glaubhaft. Legt man unter diesen Voraussetzungen mehr wahre als erfundene Aussagen zur Bewertung vor, erhöht sich der Anteil der Treffer und die falschen Alarme nehmen ab.

Die Entscheidungsausgänge werden auch durch Veränderungen im Entscheidungskriterium beeinflusst. In Abbildung und Rechenbeispiel 5.1d werden die Effekte einer liberalen Urteilsneigung dargestellt. Diese liegt vor, wenn sich Beurteiler häufiger für als gegen die Glaubhaftigkeit von Aussagen entscheiden. Der exemplarischen Darstellung wurde eine Basisrate von 50% und eine zufällige Beurteilung zugrunde gelegt. Es zeigt sich, dass eine liberale Urteilsneigung eine Erhöhung der Treffer und der falschen Alarme bewirkt.

In Abbildung und Rechenbeispiel 5.1e wurde zugleich die Basisrate auf über 50% erhöht und eine liberale Entscheidungsneigung angesetzt. Dadurch verstärkt sich die Zunahme der Treffer. Die Effekte auf die falschen Alarme gleichen sich hingegen aus.

Prozentsatz richtiger Entscheidungen

Bei dichotomen Glaubhaftigkeitsurteilen lässt sich die Urteilsgüte als Prozentsatz richtiger Entscheidungen abbilden. Richtige Entscheidungen umfassen sowohl Treffer als auch korrekte Zurückweisungen. Folglich entspricht das Mittel der Treffer- und korrekten Zurückweisungsrate der prozentualen Urteilsrichtigkeit (vgl. Rechenbeispiel 5.1a). Diese stellt den am einfachsten zu interpretierenden Index der Urteilsgüte dar. Sie hängt jedoch auch von der Basisrate ab (vgl. Park & Levine, 2001). In Laborstudien werden meist gleich viele tatsächlich wahre und erfundene Aussagen beurteilt. Dadurch liegt die zufällige

Urteilsrichtigkeit bei 50%. Für den Nachweis einer guten Beurteilungsfähigkeit sollte dieser Wert deutlich überschritten werden.

Urteilsrichtigkeit für wahre und erfundene Aussagen

Des Weiteren ist es von Interesse, ob wahre und erfundene Aussagen gleichermaßen häufig richtig beurteilt werden. Daher wird in der Täuschungsliteratur die Urteilsrichtigkeit für wahre und erfundene Aussagen getrennt betrachtet. Die Trefferrate stellt einen Indikator der Urteilsrichtigkeit für wahre Aussagen dar. Die Rate der korrekten Zurückweisungen verweist auf die Urteilsrichtigkeit für erfundene Aussagen. Statt den Anteil der Treffer an allen Entscheidungen zu ermitteln, wird lediglich ihr Anteil an den tatsächlich wahren Aussagen betrachtet. Daher impliziert eine geringe Urteilsrichtigkeit für wahre Aussagen, dass wahre Aussagen oftmals als nicht-glaubhaft beurteilt werden (falsche Zurückweisungen). Entsprechend wird der Anteil der korrekten Zurückweisungen an den erfundenen Aussagen berechnet. Die Urteilsrichtigkeit für erfundene Aussagen wird beeinträchtigt, wenn tatsächlich erfundene Aussagen als glaubhaft eingeschätzt werden (falsche Alarme).

Der Anteil der Treffer und korrekten Zurückweisungen an allen Entscheidungen wird durch Veränderungen in der Basisrate beeinflusst (vgl. Abbildung und Rechenbeispiel 5.1c). Dies wird allerdings bei der Ableitung der Urteilsrichtigkeit für wahre und erfundene Aussagen nicht beachtet. So beziehen sich die Raten der Treffer bzw. korrekten Zurückweisungen nur auf die Teilstichproben der wahren bzw. erfundenen Aussagen. Beispielsweise erhöhen sich infolge einer Basisratenverschiebung nicht nur die Treffer, sondern auch die falschen Zurückweisungen. Das Verhältnis der richtigen zu den falschen Entscheidungen für die Teilstichprobe der wahren Aussagen bleibt dadurch unverändert. Dies gilt ebenso für die Teilstichprobe der erfundenen Aussagen.

Allerdings ist zu berücksichtigen, wieviele Aussagen als glaubhaft bzw. nicht-glaubhaft beurteilt wurden. Beispielsweise könnten Beurteiler dazu neigen, Aussagen als glaubhaft einzuschätzen. Infolgedessen ergäbe sich bereits bei

zufälligen Beurteilungen für wahre Aussagen eine Urteilsrichtigkeit von mehr als 50%. Für erfundene Aussagen würde sie entsprechend geringer ausfallen. Abbildung und Rechenbeispiel 5.1d verdeutlichen, dass durch eine liberale Urteilsneigung die Trefferrate erhöht wird. Hingegen wird die Rate der korrekten Zurückweisungen reduziert.

Signalentdeckungstheoretische Indizes

Die vorangegangenen Ausführungen demonstrierten, dass die Urteilsgüte nicht nur von der Fähigkeit der Beurteiler abhängt, sondern auch von ihrer Urteilsneigung. Die Signalentdeckungstheorie differenziert explizit zwischen diesen beiden Aspekten. So werden die Urteilsgüte und -neigung getrennt voneinander quantifiziert. Die Berechnung der entsprechenden Indizes basiert auf den Raten der Treffer und falschen Alarme.

In der vorliegenden Arbeit wird die Urteilsgüte über den Diskriminationsleistungsindex A' (Rae, 1976) quantifiziert. Dieser kann Werte zwischen 0 und 1 annehmen, wobei ein Wert von .5 eine zufällige Diskriminationsleistung widerspiegelt. Zur Berechnung von A' wird je nach Verhältnis von Treffern (Hits, H) und falschen Alarmen (False Alarms, F) eine der beiden folgenden Formeln angewandt (Rae, 1976, p. 98):

Falls $H \geq F$:

$A' = (H^2 + F^2 + 3H - F - 4FH) / (4H (1-F))$,

bzw. falls $H < F$:

$A' = (H - H^2 + F - F^2) / 4F (1 - H)$.

Ob eine konkrete Aussage als glaubhaft oder nicht-glaubhaft eingeschätzt wird, hängt zudem vom Entscheidungskriterium der Beurteiler ab. Beurteiler, die eine Aussage erst dann als glaubhaft einschätzen, wenn der Glaubhaftigkeitseindruck sehr stark ausgeprägt ist, verwenden ein konservatives Entscheidungskriterium. Wenn Personen ein liberales Entscheidungskriterium ansetzen, neigen sie dazu Aussagen als glaubhaft zu beurteilen. Die Urteilsneigung B" (Donaldson, 1992) lässt sich ebenfalls auf der Grundlage der

Raten für Treffer und falsche Alarme berechnen. Dieser Index variiert zwischen -1 und 1, wobei positive Werte auf eine konservative und negative auf eine liberale Urteilsneigung verweisen. Die Berechnung erfolgt im Folgenden anhand der Formel von Donaldson (1992, p. 276):

$B'' = ((1 - H) (1 - F) - HF) / ((1-H) (1-F) + HF)$.

Welches Entscheidungskriterium sich empfiehlt, hängt von der Basisrate und den Konsequenzen der beiden Fehlerarten ab. Beispielsweise sollte ein liberales Entscheidungskriterium gewählt werden, wenn die meisten Aussagen tatsächlich wahr sind oder es besonders wichtig ist, wahre Aussagen richtig zu beurteilen. In Laborstudien wird die Basisrate in der Regel auf 50% festgelegt. Veränderungen im Entscheidungskriterium bewirken lediglich, dass ein Fehler auf Kosten des anderen minimiert wird. Erachtet man beide Fehler als gleichermaßen unerwünscht, so wäre es am sinnvollsten ein ausgewogenes Entscheidungskriterium anzusetzen. Entsprechend sollten sich Beurteiler gleichermaßen häufig für und gegen die Glaubhaftigkeit von Aussagen entscheiden. Dies würde sich in einer Urteilsneigung von $B'' = 0$ widerspiegeln.

Standardisierte Mittelwertsdifferenz d

Auch bei mehrstufiger Beurteilungsgrundlage lassen sich prozentuale Richtigkeiten und signalentdeckungstheoretische Kennwerte ermitteln. Für die Berechnung der prozentualen Urteilsrichtigkeiten werden die Urteile allerdings

meist nachträglich dichotomisiert. Dies geht mit einem Informationsverlust einher. Zudem können unentschiedene Urteile nicht verwertet werden oder müssen willkürlich den Beurteilungen als glaubhaft oder nicht-glaubhaft zugeordnet werden. Die Dichotomisierung lässt sich vermeiden, wenn man die Urteilsgüte über die standardisierte Mittelwertsdifferenz d bestimmt. Diese spiegelt die Differenz der Glaubhaftigkeitsbeurteilungen für wahre (W) und erfundene (NW) Aussagen, relativiert an deren Standardabweichungen wider (vgl. C. F. Bond & DePaulo, 2006, p. 218):

$$d = (M(W) - M(NW)) / SD_{pooled}$$

Ein Wert von d = 1 bedeutet, dass tatsächlich wahre Aussagen um eine Standardabweichung höhere Bewertungen erzielen als erfundene. Nach Cohen (1988) sind Werte von d = 0.2 als geringer, von d = 0.5 als mittlerer und von d = 0.8 als starker Effekt zu interpretieren. Das Differenzmaß zeigt allerdings nicht, wieviele Entscheidungen richtig ausgefallen sind (vgl. Kalbfleisch, 1990).

Metaanalytische Befunde

C. F. Bond und DePaulo (2006) publizierten eine Metaanalyse zur Urteilsgüte von Laien bei der Entdeckung von Täuschung. Dabei integrierten sie die Befunde von 186 englischsprachigen Studien, die bis Ende 2003 zugänglich waren. Ausgeschlossen wurden Studien, bei denen Aussagen von Kindern (Alter < 17 Jahre) hinsichtlich ihrer Glaubhaftigkeit eingeschätzt wurden, oder bei denen Hilfsmittel wie polygraphische Auswertungen oder eine inhaltsorientierte Aussageanalyse (CBCA-Merkmale) verwendet worden waren. C. F. Bond und DePaulo (2006) interessierten sich weniger für die Fähigkeit zu lügen, als für die Fähigkeit Aussagen hinsichtlich ihrer Glaubhaftigkeit richtig einzuschätzen. Daher wählten sie statt der Stimuluspersonen die Beurteiler als Analyseeinheit. Es wurden 348 unabhängige Beurteilerstichproben identifiziert, für die jeweils eine Messung der Urteilsgüte erhoben wurde. Ob diese Beurteilerstichproben dieselben oder unterschiedliche Aussagen bewerteten, wurde dabei nicht beachtet.

C. F. Bond und DePaulo (2006) leiteten verschiedene Indizes zur Bestimmung der Urteilsgüte ab. Im Folgenden werden ihre Befunde hinsichtlich der für die vorliegende Arbeit relevanten Indizes dargestellt. Dabei wird auch auf ältere quantitative Forschungsarbeiten hingewiesen (DePaulo, Zuckerman &

Rosenthal, 1980; Kalbfleisch, 1990; Kraut, 1980; Zuckerman, DePaulo & Rosenthal, 1981).[3]

Prozentsatz richtiger Entscheidungen

C. F. Bond und DePaulo (2006) fassten die Befunde von 263 unabhängigen Beurteilerstichproben zusammen, die dichotome Glaubhaftigkeitsentscheidungen trafen. Es ergab sich eine gewichtete durchschnittliche Urteilsrichtigkeit von 53.4% (95% KI = 53.1% bis 53.7%; ungewichtet: 54.0%). Die Befunde einer älteren Metaanalyse wurden damit weitestgehend repliziert. So schätzte Kraut (1980) anhand von neun Primärstudien die ungewichtete durchschnittliche Urteilsrichtigkeit mit 57.0% (SD = 7.8%) nur geringfügig höher ein. Laien sind demnach nicht besonders gut dazu in der Lage wahre und erfundene Aussagen zu unterscheiden.

Nach Kraut (1980) sind die Befunde zur Urteilsgüte aus evolutionstheoretischer Sicht nachvollziehbar. Er argumentierte, dass spezifische Verhaltensweisen, die immer bei erfundenen, jedoch niemals bei wahren Aussagen auftreten, einer natürlichen Auslese nicht Stand gehalten hätten. Schließlich würden solche eindeutigen Täuschungsindikatoren alle potentiellen Vorteile des Lügens zunichte machen. Allerdings erscheinen Fähigkeiten, die es einem erlauben, die Täuschungsversuche anderer aufzudecken, ebenfalls evolutionär vorteilhaft. Demnach sollte durch natürliche Auslese sowohl die Fähigkeit zu lügen als auch die Fähigkeit Lügen zu entdecken gefördert worden sein. Im Laufe der Evolution wäre dadurch ein Gleichgewicht zwischen beiden Fähigkeiten entstanden.

---

[3] Die metaanalytischen Befunde von Mattson, Ryan, Allen und Miller (2000) werden hingegen nicht berücksichtigt. Die Autoren beschränkten sich auf sieben Studien, die sie für die Entdeckung von Täuschung in Wirtschaftsunternehmen als relevant erachteten. Die Primärstudien waren allerdings hinsichtlich der Operationalisierung der Glaubhaftigkeitsurteile kaum vergleichbar.

Auch C. F. Bond und DePaulo (2006) erklärten die Befunde zur Urteilsgüte indem sie zwischen der Perspektive der lügenden und belogenen Person unterschieden und einen „doppelten Standard“ annahmen. Als Lügner seien Personen pragmatisch. Falschaussagen werden auf der Grundlage persönlicher Bedürfnisse und Ziele als gerechtfertigt angesehen. Allerdings wird es als moralisch verwerflich erachtet, wenn andere Personen lügen. Daher wird angenommen, dass andere Personen non- und paraverbale Anzeichen von Erregung und Schuld zeigen wenn sie lügen. Solche stereotypen Vorstellungen über das Verhalten von Lügnern sind jedoch in der Regel falsch (vgl. Studie 1). Demnach erreichen Personen nur eine geringe Urteilsgüte, weil sie unterschiedliche Maßstäbe für sich selbst und andere anwenden. Untersuchungen, in denen stereotype Täuschungsvorstellungen für das eigene Verhalten mit dem anderer kontrastiert wurden (z.B. Akehurst, Köhnken, Vrij & Bull, 1996), unterstützen diese Argumentation.

Urteilsrichtigkeit für wahre und erfundene Aussagen

C. F. Bond und DePaulo (2006) ermittelten eine gewichtete Urteilsrichtigkeit von 60.3% (95% KI = 60.1% bis 60.6%; ungewichtet: 61.5%) für tatsächlich wahre und von 48.7% (95% KI = 48.5 bis 49.0%; ungewichtet: 47.6%) für erfundene Aussagen. Laien gelingt es demnach besser tatsächlich wahre Aussagen richtig einzuschätzen, als Lügen zu entdecken. Dies lässt sich unter anderem auf ihre Urteilsneigung zurückführen. So indizierten in den von C. F. Bond und DePaulo gewichtet zusammengefassten Untersuchungen 55.0% (95% KI = 54.8% bis 55.3%; ungewichtet: 56.9%) der Urteile, dass die Aussagen den Beurteilern glaubhaft erschienen. Bei der Interpretation der Urteilsgüte für wahre bzw. erfundene Aussagen ist diesem sogenannten Wahrheitsbias (vgl. Zuckerman et al., 1981) Rechnung zu tragen. Wenn Beurteiler keinerlei Fähigkeit aufweisen wahre und erfundene Aussagen zu unterscheiden, sollten sie nach den Befunden von C. F. Bond und DePaulo 55.0% der wahren bzw. 45.0% der erfundenen zufällig richtig beurteilen. Die empirisch ermittelten Urteilsrichtigkeiten (60.3% bzw. 48.7%;

95% KIs: s.o.) fielen jeweils signifikant höher aus. Demnach werden sowohl wahre als auch erfundene Aussagen überzufällig häufig richtig beurteilt.

Kognitive Heuristiken (z.B. Levine, Park & McCornack, 1999; O'Sullivan, 2003; O'Sullivan, Ekman & Friesen, 1988; Stiff, Kim & Ramesh, 1992) und soziale Konversationsregeln (vgl. Grice, 1975; Vrij, 2008) könnten dazu beitragen, den Wahrheitsbias von Laien zu erklären. Personen werden im Alltag vermutlich wesentlich häufiger mit wahren als erfundenen Darstellungen konfrontiert (vgl. DePaulo, Kashy, Kirkendol, Wyer & Epstein, 1996). Aufgrund dieser Erfahrung könnten sie die Heuristik ableiten, dass Aussagen in der Regel wahr sind. Dies würde sich in der Tendenz manifestieren, Aussagen als glaubhaft zu beurteilen. Zudem haftet dem Begriff der Lüge etwas moralisch Verwerfliches an. Daher verbieten es soziale Konventionen jemanden als Lügner zu bezeichnen oder seine Aussage anzuzweifeln. C. F. Bond und DePaulo (2006) argumentierten, dass solche Regeln außer Kraft gesetzt werden, wenn Personen instruiert werden, die Glaubhaftigkeit von Aussagen zu beurteilen. Infolgedessen vermuten die Autoren, dass die Laborbefunde das Ausmaß des Wahrheitsbias in der Realität sogar noch unterschätzen.

Signalentdeckungstheoretische Indizes

Nur wenige Autoren verwendeten aus der Signalentdeckungstheorie abgeleitete Indizes, um die Diskriminationsleistung und Urteilsneigung getrennt voneinander zu quantifizieren (vgl. C. F. Bond & DePaulo, 2006; G. D. Bond & Lee, 2005; G. D. Bond, Malloy, Arias, Nunn & Thompson, 2005; G. D. Bond, Thompson & Malloy, 2005; Leach, Talwar, Lee, Bala & Lindsay, 2004; Meissner & Kassin, 2002; Sporer, 2004). Werden jedoch getrennte Urteilsrichtigkeiten für wahre und erfundene Aussagen berichtet, lassen sich A' und B" einfach ableiten. So entspricht die Urteilsrichtigkeit für wahre Aussagen der Trefferrate. Die Urteilsrichtigkeit für erfundene Aussagen wiederum entspricht der Rate an korrekten Zurückweisungen. Sie ist demnach direkt von der Rate der falschen Alarme abhängig (CR-Rate = 1 – F-Rate).

C. F. Bond und DePaulo (2006) verwendeten andere als die für die vorliegende Untersuchung verwendeten signalentdeckungstheoretischen Indizes. Um die Vergleichbarkeit zu ermöglichen, wurden die von C. F. Bond und DePaulo (2006) berichteten prozentualen Urteilsrichtigkeiten anhand der Formeln von Rae (1976) und Donaldson (1992) umgerechnet. Dies ergab eine Diskriminationsleistung von A' = .59 und eine Urteilsneigung von B" = -.27. Demnach verweisen auch die signalentdeckungstheoretischen Analysen auf eine nur geringfügig überzufällige Diskriminationsleistung und eine liberale Urteilsneigung.

Standardisierte Mittelwertsdifferenz d

C. F. Bond und DePaulo (2006) integrierten zunächst nur die Primärstudien, bei denen die Glaubhaftigkeitsbeurteilung anhand mehrstufiger Ratingskalen erfolgte. Über 107 Beurteilerstichproben hinweg ergab sich eine gewichtete standardisierte Mittelwertsdifferenz von d = 0.34 (95% KI = 0.31 bis 0.38; ungewichtet: d = .35). Zudem berechneten die Autoren die standardisierte Mittelwertsdifferenz für 216 Beurteilerstichproben mit dichotomer Beurteilungsgrundlage. Dazu ermittelten sie zunächst die Urteilsgüte für wahre Aussagen und den Anteil der als glaubhaft beurteilten erfundenen Aussagen. Die Differenz dieser beiden Kennwerte wurde anschließend an der Standardabweichung relativiert, die sich für die Einschätzung derselben Aussagen durch verschiedene Beurteiler ergab. Im Vergleich zu mehrstufigen Urteilen zeigte sich bei dichotomer Beurteilungsgrundlage eine signifikant höhere gewichtete standardisierte Mittelwertsdifferenz von d = 0.42 (erstes, zweites, drittes Quartil: d = 0.02, d = 0.50, d = 1.04; ungewichtet: k.A.). Bei 61 weiteren Beurteilerstichproben fehlten Angaben zu den Standardabweichungen für die Beurteilung wahrer und erfundener Aussagen. Wurden diese anhand einer Binomialverteilung geschätzt, ergab sich ein gewichtetes mittleres d von 0.40 (95% KI = k.A.; ungewichtet: k.A.). Wurden die Befunde sämtlicher 349 Beurteilerstichproben integriert, lag die gewichtete standardisierte Mittelwertsdifferenz bei d = 0.39 (95% KI = k.A.;

ungewichtet: d = 0.49). Die Urteilsgüte fiel damit deutlich geringer aus als in einer älteren Metaanalyse von DePaulo et al. (1980). Diese fanden über 16 Studien starke Effekte mit einem Median von d = 0.86. Dennoch werteten C. F. Bond und DePaulo die anhand der standardisierten Mittelwertsdifferenz quantifizierte Urteilsgüte von Laien als vielversprechend. Sie argumentierten, dass ihre Befunde der Validität objektiver Täuschungsindikatoren gegenüberzustellen sind. Eine Urteilsgüte von d = 0.40 übertrifft bei weitem die Effektstärken für 155 von 158 metaanalytisch untersuchten objektiven Indikatoren (DePaulo, Lindsay, Malone, Muhlenbruck, Charlton & Cooper, 2003). Daher schlussfolgerten C. F. Bond und DePaulo:

> “Solange die Entwicklung besserer verhaltensbasierter Technologien noch bevorsteht, sind Personen, die erfundene und wahre Aussagen am effizientesten unterscheiden wollen, besser damit beraten, naive Beurteiler um Glaubhaftigkeitsurteile zu bitten, als Zeit in die Kodierung von Verhalten zu investieren.“ (S. 51, Übersetzung der Verfasserin).

Moderatoren der Urteilsgüte

Vorbereitung und Interaktion

C. F. Bond und DePaulo (2006) fanden nur wenig Variabilität in den Ergebnissen der Primärstudien. Dennoch ließen sich Moderatoren der Urteilsgüte aufzeigen. Beispielsweise fand sich über 15 Studien, in denen das Ausmaß an Vorbereitungszeit variiert wurde, eine höhere Urteilsgüte bei unvorbereiteten (k = 42; mittleres d: k.A.) im Vergleich zu vorbereiteten Aussagen (k = 42; mittleres d: k.A.; gewichtetes mittleres d für die Differenz: d = -0.14, 95% KI = -0.21 bis -0.08) Z = 4.49, p < .01. Ebenso scheint es für die Urteilsgüte bedeutsam zu sein, ob die Stimulusperson ihre Aussage frei formuliert oder mit einer anderen Person interagiert. In keiner der von C. F. Bond und DePaulo analysierten Studien wurde das Ausmaß an Interaktion systematisch variiert. Studienübergreifende Analysen verwiesen jedoch auf eine höhere Urteilsgüte wenn die Stimuluspersonen mit

dem Beurteiler (k = 33; gewichtetes d = 0.32; ungewichtetes d = 0.23, 95% KI = 0.16 bis 0.31) oder einer anderen Person interagierten (k = 224, gewichtetes d = 0.47; ungewichtetes d = 0.42, 95% KI = 0.39 bis 0.44) im Vergleich zu freien Berichten (k = 127, gewichtetes d = 0.30; ungewichtetes d = 0.38, 95% KI = 0.34 bis 0.41), Q(2) = 57.14, p < .001.

Der sogenannte kognitive Ansatz (Zuckerman et al., 1981) bzw. dessen Erweiterung durch ein Arbeitsgedächtnismodell der Lügenproduktion (Sporer & Schwandt, 2006, 2007) könnten zur Erklärung beider Moderatoreffekte beitragen. Unterschiede zwischen wahren und erfundenen Aussagen werden auf die Belastung des Arbeitsgedächtnisses zurückgeführt. So wird angenommen, dass es mehr kognitive Anstrengung erfordert, eine Aussage frei zu erfinden, als wahrheitsgemäß zu schildern. Durch Vorbereitung könnte diese kognitive Belastung reduziert, durch die Interaktion mit einer anderen Person hingegen verstärkt werden. Infolgedessen wären auch die Unterschiede zwischen wahren und erfundenen Aussagen reduziert bzw. verstärkt. Dies würde sich wiederum auf die Schwierigkeit der Beurteilung auswirken und die vorgefundenen Unterschiede in der Urteilsgüte erklären.

Berufsgruppenunterschiede

Verschiedene Berufsgruppen werden häufig mit Situationen konfrontiert, in denen es wichtig ist, wahre und erfundene Aussagen zu unterscheiden. Daher überprüften C. F. Bond und DePaulo (2006), ob solche Berufsgruppen über eine bessere Urteilsfähigkeit verfügen als Laien. Die Autoren konnten jedoch über 16 Studien hinweg keine signifikanten Unterschiede in der Urteilsgüte feststellen. Dies entspricht den metaanalytischen Befunden von Aamodt und Custer (2006), die prozentuale Urteilsrichtigkeiten für zwölf Berufsgruppen aus insgesamt 108 Studien berichteten. Polizisten, Detektive, Richter und Psychologen (durchschnittlich 55.5%) waren ebenso wenig wie Studierende (54.2%) in der Lage zwischen wahren und erfundenen Aussagen zu unterscheiden. Lediglich eine Studie verwies auf eine höhere prozentuale Urteilsrichtigkeit von 64.1% für

Geheimdienstagenten (Ekman & O'Sullivan, 1991). Diese Studie wurde jedoch von Bond (in press[4]) wegen methodischer Probleme grundlegend kritisiert. Auch Kriminelle erzielten in einzelnen Studien eine hohe Urteilsgüte zwischen 57.8% und 65.4%, wobei dies vor allem für ältere Gefängnisinsassen galt (G. D. Bond, Thompson et al., 2005; G. D. Bond, Malloy et al., 2005; Hartwig, Granhag, Strömwall & Andersson, 2004a; vgl. auch G. D. Bond & Lee, 2005). G. D. Bond, Thompson et al. (2005) sowie G. D. Bond, Malloy et al. (2005) berichteten auch signalendeckungstheoretische Kennwerte. Allerdings verwendeten sie dabei eine andere Definition von Treffern und falschen Alarmen als die vorliegende Studie. Daher erschien es notwendig die Urteilsgüte $\underline{A'}$ anhand der Angaben zur Urteilsrichtigkeit für wahre und erfundene Aussagen abzuleiten. Für die drei Studien zur Urteilsgüte von Kriminellen ergaben sich Werte von $.64 \leq \underline{A'} \leq .77$ (G. D. Bond, Thompson et al., 2005; G. D. Bond, Malloy et al., 2005; Hartwig, et al., 2004a). Vrij und Semin (1996) fanden, dass Kriminelle weniger stereotype Täuschungsvorstellungen aufwiesen als Laien. Dies könnte ihre vergleichsweise gute Urteilsfähigkeit erklären.

Die Befunde verschiedener Untersuchungen, in denen die Urteilsgüte von Polizisten überprüft wurde, sind in Tabelle 5.2 dargestellt (vgl. Vrij & Mann, 2005, für einen narrativen Forschungsüberblick zur Entdeckung von Täuschung durch Polizisten sowie Vrij, 2008, für eine Darstellung der prozentualen Urteilsrichtigkeiten verschiedener Berufsgruppen). Um die Diskriminationsleistung und Urteilsneigung getrennt voneinander zu quantifizieren, wurden erneut signalentdeckungstheoretische Kennwerte abgeleitet.[5]

---

[4] Nach Auskunft des Herausgebers der Zeitschrift wird eine Veröffentlichung inklusive einer Replik von Ekman angestrebt.

Tabelle 5.2
Prozentuale Urteilsrichtigkeit insgesamt (Ges), für wahre (H) und erfundene Aussagen (CR), Diskriminationsfähigkeit und Urteilsneigung von Polizisten

Studie	N	S	Ges.	H	CR	A'	B"
Ekman et al. (1999)							
Verschiedene	36	10	50.8	53.9	47.8	.52	-.12
Bundesebene	23	10	73.0	66.1	80.0	.82	.34
Lokale Ebene	43	10	66.7	55.8	77.7	.76	.47
Garrido et al. (2004)[d]	121	4	47.1	26.2	68.6	.44	.72
Hartwig et al. (2006)[e,g]	41	82	56.1	57.1	55.0	.61	-.04
Kassin et al. (2005)[d,h]	57	10	48.3	63.8	32.9	.46	-.56
Köhnken (1987)[e]	20	4	47.0	63.0	32.0	.45	-.57
Mann & Vrij (2006)[a,b,f]	52	54	68.1	66.8	69.5	.77	.06
Mann et al. (2004)[a,b]	99	54	64.9	63.6	66.2	.73	.06
Mann et al. (2006)[a,b,e]	22	8	71.0	73.0	69.0	.80	-.10
Mann et al. (2007)[a,c,d]	103	14	58.0	51.0	64.0	.63	.26
Masip et al. (2003)[d]	224	2	52.0	30.6	73.3	.55	.72
Meissner & Kassin (2002)[c]	44	16	50.0	33.0	62.0	.45	.54
Porter et al. (2000)[e]	32	24	60.5	52.7	68.3	.68	.32
Vrij (1993)	91	40	49.0	51.0	46.0	.47	-.10
Vrij & Mann (2001a)[a]	65	6	64.0	70.0	57.0	.71	-.28
Vrij, Mann, Fisher et al. (2007)[e]	24	12	46.0	50.0	42.0	.43	-.16
Vrij, Mann, Kristen et al. (2007)[c,d]	68	36	50.3	62.7	38.3	.51	-.46
Vrij, Mann et al. (2006)[a,b]	37	54	72.0	70.0	73.0	.80	.07
Vrij et al. (2007)[c,d]	68	36	50.3	62.7	38.3	.51	-.46

Anm. N = Anzahl der Beurteiler; S = Anzahl der beurteilten Stimulusaussagen bzw. Aussageelemente insgesamt. [a] = es wurden reale Vernehmungsaussagen beurteilt; [b] = die Untersuchungen verwendeten dasselbe Stimulusmaterial; [c] = die Befunde wurden gemäß der für die vorliegende Arbeit verwendeten Definition von Treffern und falschen Alarmen umgerechnet; [d] = Werte für die Gesamtstichprobe der polizeilichen Beurteiler (Experimentalbedingungen wurden zusammengefasst); [e] = Werte für die Kontrollbedingung; [f] = Bedingung 1 und 3 wurden zusammengefasst; [g] = Beurteiler waren Polizeischüler. [h] = als Stimulusmaterial wurden Geständnisse von Gefängnisinsassen zu gesetzlichen Vergehen verwendet.

Für Polizisten auf Bundesebene (federal officers) berichteten Ekman, O'Sullivan und Frank (1999) eine ungewöhnlich hohe Urteilsgüte. Diese Gruppe bestand jedoch überwiegend aus Geheimdienstagenten, denen bereits eine besondere Urteilsfähigkeit attestiert wurde (Ekman & O'Sullivan, 1991; vgl. die Kritik von C. F. Bond, in press; Bond & Uysal, 2007). Des Weiteren wurden in mehreren Studien Videoaufzeichnungen realer polizeilicher Vernehmungen von Tatverdächtigen beurteilt (Mann, Vrij & Bull, 2004, 2006; Mann & Vrij, 2006; Vrij, Mann, Robbins & Robinson, 2006; Vrij & Mann, 2001a). Auch in diesen Studien erzielten Polizisten eine hohe Urteilsgüte. Die Autoren führten dies auf die Authentizität des Stimulusmaterials zurück. Allerdings wurden die verwendeten Aussagen nie einer anderen Berufsgruppe zur Beurteilung vorgelegt. Daher wäre es ebenso denkbar, dass die Aussagen besonders leicht zu bewerten waren. Schließlich fanden Studien, in denen Laien und Polizisten dasselbe Stimulusmaterial beurteilten, keine Unterschiede in der Urteilsgüte (z.B. DePaulo & Pfeifer, 1986; Meissner & Kassin, 2002; Vrij & Graham, 1997).

---

5 Studien, die keine getrennten Urteilsrichtigkeiten für wahre und erfundene Aussagen berichteten (Akehurst, Bull, Vrij & Köhnken, 2004; Ekman & O'Sullivan, 1991; Hartwig, Granhag, Strömwall & Vrij, 2004b; Vrij & Graham, 1997; Vrij & Mann, 2001b) konnten daher nicht berücksichtigt werden. DePaulo und Pfeifer (1986) berichteten zwar getrennte Urteilsrichtigkeiten, doch nur für die Gesamtstichprobe, die zu 38% aus Studierenden bestand. Bei Vrij, Akehurst, Brown und Mann (2006) wiederum finden sich nur Angaben zur Urteilsrichtigkeit für die Gesamtstichprobe von Lehrern, Sozialarbeitern und Polizisten. Schließlich wurde eine Untersuchung zur Beurteilung kindlicher Aussagen ausgeschlossen, in der andere als die hier verwendeten signalentdeckungstheoretischen Indizes für die Urteilsgüte und -neigung berichtet wurden (Leach et al., 2004).

Trotz vergleichbarer Urteilsgüte wurden wiederholt berufsbedingte Unterschiede in der Urteilsneigung festgestellt. Nach C. F. Bond und DePaulo (2006) tendieren Laien dazu Aussagen als glaubhaft zu beurteilen. Für Polizisten sind die Befunde zur Urteilsgüte vergleichsweise heterogen. Nur vereinzelt zeigte sich ein ausgeprägter Wahrheitsbias (Köhnken, 1987a; Vrij & Mann, 2001a, Vrij, Mann, Kristen et al., 2007). Dieser war auch in der Untersuchung von Kassin, Meissner und Norwick (2005) vorzufinden. Allerdings verwendeten die Autoren als Stimulusmaterial wahre und erfundene Geständnisse von Gefängnisinsassen zu gesetzlichen Vergehen. Die vorgefundene Tendenz Aussagen als glaubhaft einzuschätzen spiegelt demnach auch eine Tendenz wider die Befragten für schuldig zu halten. Die Autoren werteten ihre Befunde daher als Beleg für eine misstrauische Haltung von Polizisten. In anderen Studien wählten Polizisten ein ausgewogenes Entscheidungskriterium (Hartwig, Granhag, Strömwall & Kronkvist, 2006; Mann et al., 2004; Vrij, Mann et al., 2006; Mann & Vrij, 2006) oder tendierten zumindest im Gegensatz zu Laien nur geringfügig dazu Aussagen als glaubhaft zu beurteilen (Ekman et al., 1999; Mann, Vrij & Bull, 2006; Vrij, 1993). Wiederholt wurde jedoch auch ein Lügenbias berichtet, d.h. eine Tendenz Aussagen als nicht-glaubhaft einzuschätzen (Ekman et al., 1999; Garrido, Masip & Herrero, 2004; Mann, Vrij, Fisher & Robinson, 2007; Masip, Garrido & Herrero, 2003; Meissner & Kassin, 2002; Porter, Woodworth & Birt, 2000).

Urteilsneigungen sind nach McCornack und Levine (1990) auf relativ stabile Persönlichkeitsmerkmale der Beurteiler und auf situative Faktoren zurückzuführen. Um generelles Misstrauen als Persönlichkeitsmerkmal zu messen (Generalized Communicative Suspicion, GCS), entwickelten Levine und McCornack (1991) einen Selbstbeschreibungsfragebogen. Personen, die hohe GCS-Werte aufwiesen, tendierten eher dazu Aussagen als nicht-glaubhaft zu beurteilen, als Personen, die sich als weniger misstrauisch beschrieben (Levine & McCornack, 1991, Studie 3). Doch die Urteilsneigung ließ sich auch situativ beeinflussen. In mehreren Studien wurden Beurteiler darauf hingewiesen, dass der

Wahrheitsgehalt der zu bewertenden Aussagen zweifelhaft sei. Je deutlicher dieser Hinweis erfolgte, desto weniger Aussagen wurden als glaubhaft eingeschätzt (Levine & McCornack, 1991, Studie 3; McCornack & Levine, 1990; Millar & Millar, 1997; Toris & DePaulo, 1984). Dennoch erwies sich der Wahrheitsbias in diesen Laborstudien als recht robust. Er ließ sich zwar reduzieren, jedoch nicht vollständig aufheben. So wurden immer noch mehr als die Hälfte der beurteilungsrelevanten Aussagen als glaubhaft eingeschätzt (Levine & McCornack, 1991, Studie 3; McCornack & Levine, 1990; Millar & Millar, 1997; Toris & DePaulo, 1984).

Masip, Alonso, Garrido und Antón (2005) argumentierten, dass polizeiliche Vernehmungssituationen Misstrauen induzieren. Schließlich könnten Beschuldigte versuchen, durch Falschaussagen Bestrafungen zu vermeiden. Auch Opfer und Zeugen könnten aus Schamgefühl oder um jemanden zu schützen lügen. Wenn Polizisten wiederholt solchen Situationen ausgesetzt sind, könnte dies dazu führen, dass das zunächst situationsinduzierte Misstrauen chronisch wird. Diese Überlegungen unterstützend fanden Masip, Alonso et al. (2005), dass sich Polizisten im Vergleich zu Studierenden anhand der GCS-Skala als misstrauischer beschrieben. Dies galt jedoch nur für berufserfahrene Polizisten, nicht für Novizen. Die Autoren schlussfolgerten daher, dass Polizisten aufgrund ihrer beruflichen Tätigkeit ein relativ stabiles Misstrauen ausbilden.

Auch bei jungen Kriminellen wurde ein Lügenbias festgestellt (G. D. Bond, Malloy et al., 2005; G. D. Bond, Thompson et al., 2005; Hartwig et al., 2004a; vgl. auch G. D. Bond & Lee, 2005). Eigenen Berechnungen zufolge variierte die konservative Urteilsneigung zwischen B" = .31 und B" = .83. Zudem beschrieben sich Kriminelle im Vergleich zu nicht-inhaftierten Personen als misstrauischer anhand der GCS-Skala (G. D. Bond, Thompson et al., 2005). Analog zum beruflichen Alltag von Polizisten ist anzunehmen, dass auch diese Personengruppe oftmals Situationen ausgesetzt ist, die Misstrauen induzieren. So erscheint es für Kriminelle nicht nur wichtig selbst erfolgreich zu täuschen,

sondern auch die Lügen anderer zu erkennen (Hartwig et al., 2004a). Eine generalisierte misstrauische Haltung erscheint daher auch für diese Personengruppe nachvollziehbar. Des Weiteren fanden G. D. Bond, Malloy et al. (2005) in einer Pilotstudie, dass Gefängnisinsassen ohne Kontakt zur Außenwelt einen Lügenbias zeigten (N = 6, B'' = .75). Im Gegensatz dazu zeigten Inhaftierte, die häufig Gemeindearbeiten ausserhalb erledigten, einen Wahrheitsbias (N = 6, B'' = -.53). Daher spekulierten die Autoren, dass auch die Gefängnisumgebung Misstrauen induziert.

Subjektive Sicherheit

Des Weiteren stellt sich die Frage, ob Personen ihre Urteilsgüte richtig einschätzen. Wenn richtige Entscheidungen mit einer höheren subjektiven Sicherheit in das Urteil einhergehen, könnte letztere zusätzlichen Erkenntnisgewinn bringen. DePaulo, Charlton, Cooper, Lindsay und Muhlenbruck (1997) integrierten die Befunde von 16 Studien, die Zusammenhänge zwischen der subjektiven Sicherheit und der Urteilsgüte berichteten. Dabei ergab sich ein mittlerer gewichteter Zusammenhang von r = .04 (95% KI: .00 bis .08; ungewichtet: r = .03). Unterschiede zwischen den Ergebnissen der Primärstudien waren überwiegend auf Stichprobenfehler zurückzuführen. Die Befunde wurden von Aamodt und Custer (2006) an einer umfangreicheren Datenbasis von 58 Studien repliziert. Mit einer durchschnittlichen Effektstärke von r = .05 (CI: .02 bis .08) ergab sich zwar ein signifikanter, jedoch erneut nicht praktisch bedeutsamer Zusammenhang zwischen der Urteilsgüte und der subjektiven Sicherheit. Des Weiteren demonstrierten DePaulo et al. (1997), dass Personen ihre Urteilsfähigkeit überschätzen. In sechs Studien wurden das Glaubhaftigkeitsurteil und die subjektive Sicherheit anhand der gleichen Beurteilungsskalen erfasst. Für diese Teilstichprobe ergab sich eine mittlere Urteilsrichtigkeit von 57.2%. Die mittlere subjektive Sicherheit fiel mit 72.9% deutlich höher aus.

DePaulo et al. (1997) führten den fehlenden Zusammenhang unter anderem auf die Diskrepanz subjektiver und objektiver Indikatoren von Täuschung

zurück. So erwarten Personen beispielsweise, dass bestimmte Verhaltensweisen beim Lügen zunehmen, die tatsächlich abnehmen (vgl. Studie 1, Studie 2). Stützen sie ihre Glaubhaftigkeitsbeurteilung auf fehlerhafte Annahmen, sollte dies die Urteilsgüte beeinträchtigen. Die subjektive Sicherheit sollte hingegen zunehmen, wenn Personen Verhaltensweisen bemerken, die sie für valide Täuschungsindikatoren halten. Dadurch ließe sich erklären, dass Personen ihre Urteilsfähigkeit überschätzen. Vermutlich achten Beurteiler jedoch vereinzelt auch auf valide oder irrelevante Indikatoren. Infolgedessen wäre kein Zusammenhang zwischen der Urteilsgüte und der subjektiven Sicherheit nachzuweisen, wie es empirisch oftmals bestätigt wurde.

Trainingsstudien

Wiederholt wurde eine geringe Urteilsgüte für Laien und verschiedene Berufsgruppen festgestellt. Daher stellt sich die Frage, ob sich die Urteilsgüte durch systematisches Training verbessern lässt. Frank und Feeley (2003) integrierten die Befunde von elf Primärstudien, in denen die Leistung geschulter Beurteiler direkt mit der von Laien verglichen wurde (vgl. Vrij, 2000, für eine qualitative Analyse derselben Studien). Allerdings variierten die aus den Primärstudien abgeleiteten Effektstärken so stark ($\underline{r}$ = -.33 bis $\underline{r}$ = .59), dass eine Durchschnittsbildung wenig aussagekräftig erscheint. Die starke Variabilität ist vermutlich auf die heterogenen Trainingsinhalte zurückzuführen. Beispielsweise informierten Zuckerman, Koestner und Alton (1984) die Beurteiler lediglich über die Richtigkeit ihrer Entscheidung. Hingegen wurden die Beurteiler in anderen Studien über verschiedene non- und paraverbale Merkmale informiert (z.B. deTurck & Miller, 1990; Fiedler & Walka, 1993). Inwieweit Trainingserfolge nachweisbar waren, könnte von der Validität der trainierten Merkmale abhängig sein. Diese Argumentation wird durch eine Studie von Kassin und Fong (1999) unterstützt. Die Autoren zeigten, dass sich die Urteilsgüte verschlechtert, wenn Merkmale trainiert werden, deren Validität empirisch nicht belegt ist.

Daher erscheint es sinnvoll sich auf eine homogenere Teilstichprobe von Trainingsstudien zu beschränken. Erhalten Beurteiler lediglich Feedback zur Richtigkeit ihres Urteils, bleibt unklar, wie mögliche Leistungsverbesserungen entstehen. Aufgrund der geringen Validität non- und paraverbaler Indikatoren (vgl. DePaulo et al., 2003) erscheint es kaum sinnvoll, Beurteiler ausschließlich darüber zu informieren. Daher sind für die vorliegende Untersuchung vor allem Studien von Interesse, in denen inhaltliche Aussagemerkmale trainiert wurden. Dabei handelt es sich meist um Merkmale, welche die Hypothese eines persönlichen Erlebnisbezugs von Aussagen stützen. Solche Glaubhaftigkeitsmerkmale könnten sich besonders für die Schulung von Polizisten, die einem Lügenbias unterliegen, als nützlich erweisen (Garrido et al., 2004).

CBCA-Trainingsstudien

Im Rahmen der Criteria-Based Content Analysis (CBCA; Steller & Köhnken, 1989) wurden die wohl bekanntesten inhaltlichen Aussagemerkmale vorgestellt. Vrij (2005) gab einen tabellarischen Überblick über Studien, in denen die Urteilsgüte naiver mit denen von CBCA-trainierten Beurteilern verglichen wurde. Die zitierten Befunde sind jedoch nicht direkt vergleichbar. Beispielsweise beziehen sich die Angaben von Landry und Brigham (1992) auf das subjektive Glaubhaftigkeitsurteil der Beurteiler. Hingegen indiziert der von Ruby und Brigham (1998) übernommene Wert, wie gut sich der Wahrheitsstatus von Aussagen durch die Beurteilung von 15 CBCA-Merkmalen diskriminanzanalytisch vorhersagen lässt.

Es wurden nur wenige Studien publiziert, in denen die subjektiven Glaubhaftigkeitsurteile trainierter Beurteiler mit denen von Laien verglichen wurden. Die bisherigen Befunde sind in Tabelle 5.3 aufgeführt.[6] Meist wurden CBCA-Trainingseffekte anhand eines Between-Subjects-Designs untersucht (Köhnken, 1987a; Krahé & Kundrotas, 1992; Landry & Brigham, 1992; Ruby & Brigham, 1998; Steller, 1989). Akehurst et al. (2004) verwendeten hingegen ein Within-Subjects-Design.

Köhnken (1987a) überprüfte die Trainierbarkeit von Polizisten. Das zu beurteilende Stimulusmaterial umfasste Aussagen zu sozialen Ereignissen, die den Stimuluspersonen zuvor auf Video präsentiert worden waren. Zwei Personen gaben diese Ereignisse wahrheitsgemäß, zwei weitere verfälscht wieder. Alle Stimuluspersonen hatten fünf Minuten Zeit, um ihre Aussage vorzubereiten. Die insgesamt 80 Beurteiler wurden einer von vier Bedingungen zugeordnet. Die Kontrollgruppe erhielt eine allgemeine Einweisung von 15 Minuten. Die drei übrigen Gruppen wurden in 60-minütigen Schulungen über nonverbale, paraverbale oder fünf CBCA-Merkmale informiert. Den Beurteilern wurde mitgeteilt, dass Veränderungen im jeweiligen Verhaltensbereich auf verfälschte Darstellungen hinweisen könnten. Als Vergleichsbasis sahen sie zunächst wahrheitsgemäße Schilderungen der jeweiligen Stimulusperson. Die nachfolgenden beurteilungsrelevanten Aussagen wurden von allen Beurteilern

---

[6] Die Untersuchungen von Tye, Amato, Honts, Devitt und Peters (1999) sowie von Santilla, Roppola, Runtti und Niemi (2000) wurden dabei nicht berücksichtigt. In beiden Studien wurde die Güte subjektiver Glaubhaftigkeitsurteile mit der regressionsanalytisch bzw. diskriminanzanalytisch ermittelten Klassifikationsgüte von CBCA-Merkmalen verglichen. Im Gegensatz dazu bezieht sich die vorliegende Zusammenfassung ausschließlich auf Untersuchungen, in denen die subjektiven Glaubhaftigkeitsurteile verschiedener Beurteilergruppen analysiert wurden.

Tabelle 5.3
Prozentuale Urteilsrichtigkeiten für insgesamt (Ges), für wahre (H) und erfundene Aussagen (CR) sowie Diskriminationsleistung und Urteilsneigung für naive und CBCA-trainierte Beurteiler

	Kontrollgruppe							Trainingsgruppe						
	n	S	Ges	H	CR	A'	B''	n	S	Ges	H	CR	A'	B''
Akehurst et al., 2004	58	49	66.7	a	a	b	b	58	49	63.3	a	a	b	b
Köhnken, 1987a	20	4	47.0	63.0	32.0	.45	-.57	20	4	48.0	69.0	26.0	.44	-.73
Krahé & Kundrotas, 1992	22	30	63.3	a	a	b	b	31	30	74.2	a	a	b	b
Landry & Brigham, 1992	50	12	46.9	58.6	35.2	.44	-.45	64	12	55.3	75.4	35.0	.61	-.70
Ruby & Brigham, 1998	26	12	35.3	a	a	b	b	26	12	46.9	a	a	b	b
Steller, 1989	25	176	60.0	68.0	47.0	.63	-.41	3	176	71.9	77.7	62.3	.79	-.36

Anm. n = Anzahl der Beurteiler; S = Anzahl der von der gesamten Untersuchungsgruppe beurteilten Stimulusaussagen; a = Werte wurden nicht berichtet; b = Kennwerte ließen sich aufgrund fehlender Angaben nicht berechnen.

entweder als glaubhaft oder als nicht-glaubhaft eingeschätzt. Keine der drei Trainingsgruppen (nonverbal: 42.0%; paraverbal: 40.0%; CBCA: 48.0%) erzielte eine bessere Leistung als die Kontrollgruppe (47.0%). Die Urteilsgüte fiel in dieser Studie allerdings insgesamt sehr niedrig aus. Zudem wurde nicht überprüft, ob die in den Schulungen erläuterten Merkmale in dem verwendeten Stimulusmaterial vorzufinden waren. Allen Stimuluspersonen fehlte ein persönlicher Erlebnisbezug, da sie die zu schildernden Ereignisse nur beobachtet hatten. Zudem war für die verfälschten Aussagen lediglich ein Detail abzuändern. Dies dürfte einfacher sein, als eine Falschaussage frei zu erfinden. Daher wäre es möglich, dass wahre und verfälschte Aussagen tatsächlich kaum zu unterscheiden waren.

Steller (1989) fand positive Effekte eines CBCA-Trainings. Er überprüfte die Urteilsgüte anhand des Stimulusmaterials einer Untersuchung von Steller, Wellershaus und Wolf (1988, zitiert nach Steller, 1998; vgl. Steller, Wellershaus & Wolf, 1992). Dabei berichteten Kinder über negative persönliche Ereignisse wie medizinische Behandlungen, körperliche Angriffe durch andere Kinder oder Tiere. Jedes Kind formulierte sowohl eine wahre als auch eine erfundene Aussage. Steller ließ die transkribierten Aussagen von einer Trainingsgruppe und einer naiven Kontrollgruppe beurteilen. Das Training bestand aus einer 90-minütigen Informations- und Übungsveranstaltung zu 18 CBCA-Merkmalen. Die trainierten Beurteiler schätzten für jeweils 194 Aussagen das Vorhandensein der einzelnen CBCA-Merkmale anhand 4-stufiger Skalen ein und gaben anschließend ein Glaubhaftigkeitsurteil ab. Die naive Vergleichsgruppe bearbeitete insgesamt 176 der 194 Aussagen, wobei jeder Beurteiler nur die Glaubhaftigkeit von jeweils 40 Aussagen einschätzte. Die Glaubhaftigkeitsbeurteilung erfolgte in beiden Gruppen anhand 5-stufiger Skalen und unentschiedene Urteile (4.5% für die Gruppe der naiven; 9.5% für die Gruppe der trainierten Beurteiler) wurden von den Analysen ausgeschlossen. Über 176 Aussagen erzielte die Trainingsgruppe eine höhere durchschnittliche Urteilsrichtigkeit (71.9%) als die naive Kontrollgruppe (60.0%).

Die CBCA-informierten Beurteiler waren sowohl bei der Einschätzung wahrer als auch erfundener Aussagen erfolgreicher als die naiven.

Krahé und Kundrotas (1992) ließen 30 Vernehmungsprotokolle zu Sexualdelikten beurteilen. Demnach wurden nicht die Aussagen der vermeintlichen Opferzeugen selbst, sondern lediglich deren Wiedergabe durch Vernehmungsbeamte bewertet. Die Hälfte der Anzeigen war aufgrund von Tätergeständnissen oder Sachbeweisen an die Staatsanwaltschaft weitergeleitet worden. Diese Beschuldigungen wurden daher als wahr aufgefasst. Die andere Hälfte der Anzeigen wurde zurückgezogen und entsprechend als erfunden erachtet. Insgesamt 30 Polizisten wurden in einer 90-minütigen Schulung über sämtliche 19 CBCA-Merkmale informiert. Alle beurteilten danach jeweils vier der insgesamt 30 Vernehmungsprotokolle. Zunächst überprüften die Beurteiler die CBCA-Merkmale, und gaben dann ein dichotomes Glaubhaftigkeitsurteil ab.

Das subjektive Glaubhaftigkeitsurteil fiel in 74.2% der Fälle richtig aus. Eine Kontrollgruppe von 22 ungeschulten Polizisten erzielte eine geringere, jedoch im Vergleich zu anderen Studien ebenfalls ungewöhnlich hohe prozentuale Urteilsrichtigkeit von 63.3%. Dies verweist auf kritische Aspekte hinsichtlich des verwendeten Stimulusmaterials. Vernehmungsprotokolle sind vermutlich nicht frei von den Wertungen des Protokollanten. Diese könnten sich in der Auswahl und der Art der Darstellung fallbezogener Äußerungen widerspiegeln. Zudem bleibt unklar, ob die Protokollanten über die CBCA-Merkmale informiert waren. Infolgedessen kann nicht ausgeschlossen werden, dass sich ihre Darstellung daran orientierte. Auch lässt sich der objektive Wahrheitsstatus der Aussagen durch die verwendeten Außenkriterien nicht zweifelsfrei belegen. Unter anderem wäre es denkbar, dass eine Kriteriumskontamination vorlag. Möglicherweise zogen Opferzeugen ihre Anzeige zurück, weil es ihnen nicht gelang, den Vernehmungsbeamten von der Glaubhaftigkeit ihrer Aussage zu überzeugen. Eine CBCA-Analyse erfordert, dass die Originalaussagen der Stimuluspersonen vollständig und wörtlich vorliegen. Protokollierungen durch Dritte sind

grundsätzlich selektiv und bieten daher keine geeignete Basis für die Anwendung der CBCA-Merkmale. Infolgedessen lässt sich die Untersuchung von Krahé und Kundrotas (1992) nicht als Beleg für die Effektivität eines CBCA-Trainings werten.

Landry und Brigham (1992) berichteten ermutigende Befunde zur Trainierbarkeit von Studierenden. Sie verwendeten kurze Aussagen von Erwachsenen zu persönlich relevanten Lebensereignissen als Stimulusmaterial. Jede Person berichtete nach einer zweitägigen Vorbereitungszeit zunächst ein erfundenes und dann ein erlebnisbasiertes Ereignis. Für die Beurteilung wurde variiert, ob diese Aussagen über Video oder schriftlich dargeboten wurden. Des Weiteren schätzte ein Teil der Beurteiler die Glaubhaftigkeit der Aussagen naiv ein, während andere zuvor über 14 CBCA-Merkmale informiert wurden. In einem 45-minütigen Training wurde darauf hingewiesen, dass mehr als fünf CBCA Kriterien ein guter Indikator für die Glaubhaftigkeit von Aussagen seien. Die trainierten Beurteiler schätzten zunächst die Präsenz der 14 CBCA-Merkmale ein und urteilten dann über die Glaubhaftigkeit der Aussagen. Die Glaubhaftigkeitsbeurteilung erfolgte in beiden Gruppen anhand 9-stufiger Skalen. Beurteilungen, die sich nicht eindeutig als Entscheidungen für bzw. gegen die Glaubhaftigkeit der Aussagen interpretieren ließen (10.1% der Urteile), wurden von den Analysen ausgeschlossen.

Beiden Beurteilergruppen gelang es besser tatsächlich wahre im Vergleich zu erfundenen Aussagen richtig einzuschätzen. Die trainierten Beurteiler entschieden sich jedoch häufiger für die Glaubhaftigkeit der Aussagen (61.8%) als die naiven (55.2%). Dadurch waren die trainierten Beurteiler bei der Bewertung wahrer Aussagen überlegen und erzielten auch insgesamt eine signifikant höhere Urteilsrichtigkeit als die naiven. Zudem ergab sich für Videoaufzeichnungen (55.0%) eine höhere Urteilsrichtigkeit als für Transkripte (47.2%). Eine überzufällige Urteilsrichtigkeit (58.1%) wies nur die Gruppe auf, die trainiert wurde und Videoaufzeichnungen beurteilte.

Später verwendeten Ruby und Brigham (1998) ähnliche Durchführungsmodalitäten. Das Stimulusmaterial war mit dem von Landry und Brigham (1992) verwendeten vergleichbar. Ruby und Brigham untersuchten jedoch 15 CBCA-Merkmale und veränderten die Entscheidungsregel geringfügig. So wurden die Beurteiler darauf aufmerksam gemacht, dass mindestens fünf Merkmale ein guter Glaubhaftigkeitsindikator seien. Allerdings zeigten nachträgliche Analysen, dass nur 55.0% der Urteile mit dieser Entscheidungsregel konsistent waren. Ruby und Brigham (1998) ließen nur schriftliche Transkripte beurteilen, variierten jedoch die ethnische Gruppenzugehörigkeit der Stimuluspersonen. Es konnten unabhängig von der Ethnie der Stimulusperson keine Trainingseffekte hinsichtlich der Glaubhaftigkeitsbeurteilung nachgewiesen werden. Es zeigten sich jedoch Unterschiede in der Häufigkeit, mit der die CBCA-Merkmale in den Transkripten vorzufinden waren. So erzielten zehn Merkmale höhere Beurteilungen, wenn es sich um Aussagen von Personen afroamerikanischer Herkunft handelte, während zwei Merkmale bei weissen Stimuluspersonen stärker ausgeprägt waren. Allerdings wurde insgesamt auch die Validität der CBCA-Merkmale für das verwendete Stimulusmaterial nicht belegt. Das CBCA-Training ging von einer stärkeren Ausprägung der inhaltlichen Glaubhaftigkeitsmerkmale bei wahren als bei erfundenen Aussagen aus. Dies konnte zwar für sieben Merkmale bestätigt werden, doch die fünf Merkmale logische Struktur, kontextuelle Einbettung, indirekt handlungsbezogene Schilderungen, Schilderung eigener psychischer Vorgänge und Selbstbelastungen waren entgegen der Erwartung bei erfundenen Aussagen stärker ausgeprägt. Daher erscheint es folgerichtig, dass positive Trainingseffekte hinsichtlich der subjektiven Glaubhaftigkeitsurteile ausblieben.

Unklar bleibt jedoch, warum die CBCA-Merkmale nicht erwartungsgemäß zwischen wahren und falschen Aussagen differenzierten. Sieben der von Ruby und Brigham (1998) verwendeten Aussagen wurden bereits von Landry und Brigham (1992) erhoben, auch wenn sie in deren Untersuchung nicht genutzt wurden. Für

die übrigen Aussagen wurden identische Durchführungsmodalitäten verwendet. Bereits Landry und Brigham hatten Schwierigkeiten, die Validität des CBCA-Merkmals logische Struktur nachzuweisen, das häufiger bei erfundenen Aussagen vorzufinden war. Zudem erzielten wahre und erfundene Aussagen vergleichbare Bewertungen für das Merkmal Selbstbelastungen. Dennoch konnten Landry und Brigham die Validität für 10 CBCA-Merkmale empirisch nachweisen und nur für zwei erwartungskonträre Befunde feststellen. Um die Ausprägung der CBCA-Merkmale zu bestimmen, fassten Landry und Brigham die Einschätzungen von Beurteilern zusammen, von denen jeweils die Hälfte Transkripte bzw. Videoaufzeichnungen analysiert hatte. Im Gegensatz dazu basierten die CBCA-Einschätzungen in der Studie von Ruby und Brigham ausschließlich auf der Analyse schriftlicher Aussagen. Daher argumentierten Ruby und Brigham, dass CBCA-Beurteilungen möglicherweise eher valide sind, wenn auch non- und paraverbale Informationen berücksichtigt werden können. Die Autoren weisen jedoch auch darauf hin, dass die CBCA für die Analyse schriftlicher Aussagen entwickelt wurde. Ihre Validität sollte daher auch bei Transkripten nachweisbar sein.

Akehurst et al. (2004) untersuchten die Effektivität eines CBCA-Trainings anhand eines Within-Subjects-Designs. Als Stimulusmaterial dienten Aussagen von Kindern zu einer selbsterlebten, beobachteten oder erfundenen Interaktion mit einem Photographen. Studierende, Polizisten und Sozialarbeiter beurteilten zunächst jeweils vier Transkripte naiv. Dann durchliefen sie ein vierstündiges Training zu 13 CBCA-Merkmalen, deren Beurteilung sie einüben konnten. Anschließend bearbeiteten sie erneut jeweils vier Transkripte. Dabei schätzten sie die Ausprägung der CBCA-Merkmale ein und gaben ein dichotomes Glaubhaftigkeitsurteil ab. Vor dem Training zeigte sich eine ebenso gute Urteilsrichtigkeit (66.7%) wie nach dem Training (63.3%). Insgesamt war also kein Trainingseffekt festzustellen. Getrennte Analysen der drei Beurteilergruppen

zeigten jedoch, dass sich die Leistung der Polizisten infolge des Trainings verschlechterte (vorher: 68.0%; nachher: 53.0%).

In dieser wie in anderen Trainingsstudien wurde leider darauf verzichtet, die prozentuale Urteilsrichtigkeit für wahre und erfundene Aussagen getrennt zu berichten. Infolgedessen ließen sich keine signalentdeckungstheoretischen Kennwerte zur Urteilsgüte und -neigung ableiten. Durch ein CBCA-Training wurden Beurteiler auf Merkmale hingewiesen, welche die Hypothese eines persönlichen Erfahrungsbezugs stärken. Daher ist zu erwarten, dass die Tendenz Aussagen als glaubhaft zu beurteilen verstärkt wird. Die Befunde von Köhnken (1987a) sowie von Landry und Brigham (1992) unterstützen diese Argumentation.

Zudem wäre es denkbar, dass sich ein CBCA-Training nicht nur auf die Glaubhaftigkeitsurteile auswirkt, sondern beispielsweise auch auf die subjektive Sicherheit der Beurteiler. In zwei der aufgeführten Trainingsstudien wurde die subjektive Sicherheit der Beurteiler unabhängig vom Glaubhaftigkeitsurteil erhoben (Akehurst et al., 2004; Köhnken, 1987a). Es wurden jedoch keine Einflüsse des Trainings auf die subjektive Sicherheit festgestellt.

Zusammenfassend ergaben Trainingsstudien sowohl positive (Krahé & Kundrotas, 1992; Landry & Brigham, 1992; Steller, 1989) als auch ernüchternde Forschungsbefunde (Akehurst et al., 2004; Köhnken, 1987a; Ruby & Brigham, 1998). Diese Widersprüche lassen sich durch studienübergreifende Vergleiche nicht klären. So basieren nach Vrij (2005) Glaubhaftigkeitsbeurteilungen von Laien oftmals auf Videoaufzeichnungen. Hingegen analysieren CBCA-Beurteiler in der Regel schriftliche Transkripte. Infolgedessen könnten Unterschiede in der Urteilsgüte von Laien und CBCA-trainierten Beurteilern ebenso gut auf die Darbietung des Stimulusmaterials und den Umfang der verwendeten Aussagen wie auf ihren Wissensstand zurückzuführen sein.

Metaanalytisch fand Kalbfleisch (1990) keine Unterschiede in der Urteilsgüte bei audiovisueller (58.0%) und schriftlicher Darbietung von Aussagen (60.0%). Die Urteilsgüte fiel allerdings geringer aus, wenn ausschließlich visuelle

Informationen dargeboten wurden (52.0%). Diese Befunde wurden von C. F. Bond und DePaulo (2006) weitestgehend repliziert. So ergab sich eine geringere Urteilsgüte für visuell im Vergleich zu audiovisuell präsentierten Aussagen (gewichtetes mittleres d = -.44, Z = -15.72, p < .0001). Zwischen der Urteilsgüte für Transkripte und audiovisueller Darbietung war hingegen kein Unterschied festzustellen (k.A.). Für spezifisches Stimulusmaterial könnte die Urteilsgüte bei schriftlicher und audiovisueller Darbietung durchaus variieren (vgl. Landry & Brigham, 1992). Daher sind weitere Trainingsstudien erforderlich, um den widersprüchlichen Forschungsstand und offene Forschungsfragen zu klären.

Trainingsstudien mit dem Realitätsüberwachungsansatz

Auch aus dem Realitätsüberwachungsansatz (RÜ, Johnson & Raye, 1981) wurden inhaltliche Merkmale zur Unterscheidung wahrer und erfundener Aussagen abgeleitet. Die Validität dieser Merkmale wurde wiederholt untersucht (vgl. Masip, Sporer, Garrido & Herrero, 2005; Sporer, 2004; für Zusammenfassungen). RÜ-Merkmale erlaubten es mindestens genauso gut wie CBCA-Merkmale wahre und erfundene Aussagen richtig zu klassifizieren (Granhag, Strömwall & Landström, 2006; Sporer, 1997a, 1997b; Strömwall, Bengtsson, Leander & Granhag, 2004; Vrij, Akehurst, Soukara & Bull, 2004a; Vrij, Akehurst, Soukara & Bull, 2004b; Vrij, Edward & Bull, 2001b; Vrij, Edward, Roberts & Bull, 2000). Zudem demonstrierte Sporer (1997b), dass die subjektiven Glaubhaftigkeitsurteile eigenständig zur diskriminanzanalytischen Klassifikationsgüte von RM- und CBCA-Merkmalen beitrugen.

Allerdings liegen unseres Wissens keine Untersuchungen vor, in denen die subjektiven Glaubhaftigkeitsurteile RÜ-geschulter und naiver Beurteiler verglichen wurden. Insbesondere vor dem Hintergrund, dass die RÜ-Merkmale leicht zu trainieren sind (Sporer, 1997a; vgl. auch Küpper & Sporer, 1995), ist dies überraschend. Sporer und Bursch (1996) untersuchten jedoch die Auswirkungen eines Trainings, das sowohl Informationen zu CBCA- als auch zu RÜ-Merkmalen vermittelte. Dazu wählten sie ein Within-Subjects-Design. Zunächst schätzten drei

Beurteiler die Glaubhaftigkeit von 200 Aussagen zu persönlich bedeutsamen Lebensereignissen naiv ein. Danach wurden die Beurteiler über 13 CBCA- und acht RÜ-Merkmale informiert. Sie analysierten dasselbe Stimulusmaterial anhand dieser Merkmale und gaben erneut Glaubhaftigkeitsurteile ab. Vor und nach dem Training verwendeten sie dafür 10-stufige Beurteilungsskalen, die anschließend dichotomisiert wurden. Die Urteilsgüte der drei Beurteiler lag vor dem Training nur geringfügig über dem Zufallsniveau (A' = .60). Nach dem Training ergab sich eine etwas höhere Urteilsgüte (A' = .69). Allerdings zeigte sich auch ein leichter Wahrheitsbias (B" = -.14), der vor dem Training (B" = .01) nicht festzustellen war. Die Einflüsse des Wissens um CBCA- und RÜ-Merkmale auf die Urteilsgüte getrennt abzuschätzen, ist jedoch nicht möglich. Dazu wären weitere Trainingsstudien erforderlich.

ARJS-Trainingsstudien

Die Aberdeen Report Judgement Scales (ARJS; Sporer, 1996/1998/2004) umfassen 52 theoretisch abgeleitete Glaubhaftigkeitsmerkmale. Bei ihrer Entwicklung wurden sowohl CBCA- und RÜ-Merkmale als auch sozialpsychologische und gedächtnispsychologische Theorien und Forschungsbefunde berücksichtigt (vgl. Studie 3). Die Güte der subjektiven Glaubhaftigkeitsurteile ARJS-informierter Beurteiler wurde wiederholt untersucht (Barnier, Sharman, McKay & Sporer, 2005; Sporer, 1998; Sporer & Burghardt, 2004; Sporer, Samweber & Stucke, 2000; Sporer & Walther, 2006). Zudem wurde in drei Studien die Urteilsgüte für ARJS-trainierte und naive Beurteiler verglichen. Die Befunde sind in Tabelle 5.4 aufgeführt.

Sporer (1998) überprüfte erstmals, ob Beurteiler von ihrem Wissen um diese ARJS-Merkmale profitieren. Zwei Beurteiler wurden kurz in der Anwendung der ARJS geschult. Sie analysierten daraufhin wahre und falsche Aussagen zu einer nächtlichen Militärübung. Zudem schätzten sie die Glaubhaftigkeit jeder Aussage intuitiv anhand 10-stufiger Skalen ein. Ihre Urteile wurden später gemittelt und dichotomisiert.

Tabelle 5.4

Prozentuale Urteilsrichtigkeiten insgesamt (Ges), für wahre (H) und erfundene Aussagen (CR) sowie Diskriminationsleistung und Urteilsneigung für naive und ARJS-trainierte Beurteiler

	Kontrollgruppe							Trainingsgruppe						
Studie	n	S	Ges	H	CR	A’	B"	n	S	Ges	H	CR	A’	B"
Sporer, 1998	2	71	58.4	62.5	54.2	.64	-.17	2	71	69.0	75.0	62.9	.78	-.28
Sporer et al., 2002	54	144	57.6	66.7	48.6	.64	-.36	54	144	62.5	70.8	54.2	.70	-.35
Sporer & Masip, 2008[x]	32	64	51.4	60.5	42.2	.53	-.35	32	64	50.6	59.4	41.8	.51	-.34
Sporer & Masip, 2008[y]								29	64	55.0	53.0	56.9	.59	.08

Anm. n = Anzahl der Beurteiler; S = Anzahl der von der gesamten Untersuchungsgruppe beurteilten Stimulusaussagen. [x] = Kurzanleitung über 17 ARJS-STV-S-Merkmale, [y] = Kurzanleitung über 4 ARJS-STV-S-Merkmale.

Dabei wurden Werte zwischen 1 und 5 einer Beurteilung als nicht-glaubhaft zugeordnet und Werte von 5.5 bis 10 einer Beurteilung als glaubhaft. Die Richtigkeit der subjektiven Glaubhaftigkeitsurteile variierte in Abhängigkeit von der Vorbereitung der Stimuluspersonen. Die ARJS-Beurteiler schätzten unvorbereitete Aussagen (80.6%) häufiger richtig ein als vorbereitete (57.1%). Insgesamt erzielten die beiden ARJS-Beurteiler eine höhere Urteilsgüte (69.0%) als zwei Personen, die dieselben Aussagen naiv beurteilten (58.4%). Allerdings zeigten die ARJS-geschulten Beurteiler auch einen stärkeren Wahrheitsbias als die naiven.

Sporer et al. (2000) führten eine umfangreichere Trainingsuntersuchung durch. Als Stimulusmaterial wurden 72 wahre, 36 verfälschte und 36 frei erfundene Aussagen zu einem Restaurantbesuch verwendet. Insgesamt 108 Beurteiler wurden einer von zwei Gruppen zugewiesen. Die Hälfte der Beurteiler erhielt schriftliche Kurzdefinitionen und Beschreibungen für neun ARJS-Merkmale. Sie bewerteten zunächst jedes ARJS-Merkmal und gaben dann ein 10-stufiges Glaubhaftigkeitsurteil ab. Die Vergleichsgruppe der naiven Beurteiler schätzte nur die Glaubhaftigkeit der Aussagen anhand 10-stufiger Skalen ein. Alle Beurteiler bewerteten jeweils 16 der insgesamt 144 transkribierten Aussagen. Die ARJS-informierten Beurteiler erzielten eine etwas höhere Urteilsgüte (62.5%) als die naiven (57.6%). Die Tendenz Aussagen als glaubhaft zu bewerten war in beiden Gruppen gleichermaßen ausgeprägt.

Sporer und Masip (2007) nutzten erstmals die Kurzform der ARJS (ARJS--Short Training Version--Spanish, ARJS-STV-S), um Beurteiler anzuleiten. Dieselben 64 transkribierten Aussagen zu moralischen und gesetzlichen Vergehen wurden von drei Beurteilergruppen bearbeitet. Die erste Gruppe erhielt eine schriftliche Kurzanleitung zu sämtlichen 17 Merkmalen der ARJS-STV-S. Die Beurteilung der einzelnen Merkmale sollte anhand vorgegebener Regeln zu einem Glaubhaftigkeitsurteil integriert werden. Eine zweite Gruppe wurde lediglich über vier ARJS-STV-S-Merkmale informiert. Diese Merkmale hatten für das verwendete Stimulusmaterial die höchste Validität erzielt. Auch dieser Gruppe wurden

Gewichtungsregeln zur Ableitung des Glaubhaftigkeitsurteils vorgegeben. Schließlich wurden subjektive Glaubhaftigkeitsbeurteilungen einer naiven Kontrollgruppe erhoben. Jeder Beurteiler bewertete jeweils 16 Aussagen anhand 10-stufiger Skalen. Über alle Aussagen hinweg war die Urteilsgüte der drei Beurteilergruppen vergleichbar. Es zeigte sich jedoch eine Wechselwirkung zwischen dem Wahrheitsstatus und der Gruppenzugehörigkeit auf die Urteilsrichtigkeit. Bei der Beurteilung wahrer Aussagen waren keine Unterschiede zwischen den drei Gruppen festzustellen. Bei der Beurteilung erfundener Aussagen erzielte jedoch die Gruppe, die anhand der vier validen ARJS-STV-S-Merkmale angeleitet worden war, mit 56.9% die höchste Urteilsrichtigkeit. Sie zeigte damit eine signifikant bessere Leistung beim Entdecken von Lügen als die Kontrollgruppe (42.2%) und die über 17 ARJS-Merkmale informierten Beurteiler (41.8%). Zudem zeigten sich Unterschiede in der Urteilsneigung. Die über vier ARJS-STV-S-Merkmale informierte Beurteilergruppe setzte als einzige ein ausgewogenes Entscheidungskriterium an. Die beiden anderen Gruppen tendierten dazu, Aussagen als glaubhaft einzuschätzen.

Zusammenfassend konnten bereits positive Effekte einer ARJS-Schulung auf die Urteilsgüte nachgewiesen werden (Sporer, 1998). Allerdings wurde bislang nicht belegt, dass Beurteiler bei wahren Aussagen von ihrem Wissen um die Merkmale der Kurzform der ARJS profitieren (vgl. Sporer & Masip, 2007). Hinsichtlich der Urteilsneigung wurden widersprüchliche Ergebnisse erzielt, die es durch weitere Untersuchungen zu klären gilt. Ebenso fehlen Studien, die mögliche Effekte einer ARJS-Schulung auf die subjektive Urteilssicherheit überprüfen.

Brunswiksche Linsenmodellanalyse

Die meisten Trainingsstudien analysierten die Validität der inhaltlichen Aussagemerkmale und die Güte der subjektiven Glaubhaftigkeitsurteile getrennt. Trainingseffekte hängen jedoch sowohl von der Validität der Aussagemerkmale ab

als auch von ihrem Einfluss auf das subjektive Glaubhaftigkeitsurteil. Daher verwendeten einzelne Autoren das Brunswiksche Linsenmodell (vgl. Hammond & Stewart, 2001), um die Produktion und Beurteilung wahrer und erfundener Aussagen zu untersuchen (z.B. Fiedler, 1989a; Fiedler & Walka, 1993; Köhnken, 1990; Reinhard, Burghardt, Sporer & Bursch, 2002; Sporer, 1998; Sporer & Küpper, 1995). Dabei werden die Aussagemerkmale einerseits mit dem objektiven Wahrheitsstatus in Beziehung gesetzt. Daraus ergeben sich die sogenannten ökologischen Validitäten der Merkmale. Andererseits werden die Zusammenhänge zwischen den inhaltlichen Merkmalen und den subjektiven Glaubhaftigkeitsurteilen untersucht. Dies zeigt wie Beurteiler die Merkmale nutzen, um zu einem Urteil zu gelangen.

Fiedler (1989a, vgl. auch Fiedler, 1989b, Studie 2) ließ 32 wahre und 32 erfundene Aussagen von zwölf Beurteilern hinsichtlich ihrer Glaubhaftigkeit bewerten. Ein Teil der Beurteiler analysierte dieselben Aussagen zusätzlich hinsichtlich sechs vorgegebener Merkmale. Diese wurden aus naiven Theorien der Lüge abgeleitet. Einerseits wurden inhaltliche (sozial Unerwünschtes, Seltenheit und Ausgefallenheit) und nonverbale Merkmale (Auffälligkeiten in der Körpersprache) vorgegeben. Doch auch allgemeine Eindrücke (verdächtige Persönlichkeit, Intimität) und die Verifizierbarkeit von Aussagen (faktische Aussagen mit einem eindeutigen Wahrheitskriterium versus Aussagen über Meinungen) wurden untersucht. Die Auswertung erfolgte über zwei multiple Regressionsanalysen, bei denen die sechs Merkmale als Prädiktoren dienten. Als Kriterium wurde zum einen der objektive Wahrheitsstatus der Aussagen, zum anderen das subjektive Glaubhaftigkeitsurteil herangezogen. Es zeigte sich, dass die Merkmale stark mit dem subjektiven Glaubhaftigkeitsurteil zusammenhingen ($\underline{R}$ = .69). Die ökologischen Validitäten fielen hingegen geringer aus ($\underline{R}$ = .33). Entsprechend korrelierten der objektive Wahrheitsstatus und das subjektive Glaubhaftigkeitsurteil nur gering ($\underline{r}(62)$ = .19). Die Beurteiler orientierten sich zu

sehr an Merkmalen, die nicht valide zwischen wahren und erfundenen Aussagen unterschieden.

Sporer und Küpper (1995) analysierten die Glaubhaftigkeitsurteile eines Beurteilers, der zunächst die inhaltliche Qualität des Stimulusmaterials eingeschätzt hatte. Insgesamt stellten 100 Versuchspersonen je ein erlebnisbasiertes und ein erfundenes Lebensereignis schriftlich dar. Ein Beurteiler bewertete sämtliche Darstellungen anhand einer adaptierten Fassung des Memory Characteristics Questionnaires (Johnson, Foley, Suengas & Raye, 1988). Diese Bewertungen ließen sich faktorenanalytisch auf acht RÜ-Merkmale reduzieren. Zudem schätzte der Beurteiler die Glaubhaftigkeit jedes schriftlich beschriebenen Ereignisses anhand 10-stufiger Skalen ein. Es ergab sich ein vergleichsweise starker Zusammenhang zwischen dem objektiven Wahrheitsstatus und dem subjektiven Glaubhaftigkeitsurteil ($r(198) = .45$). Dennoch wiesen die RÜ-Merkmale stärkere Zusammenhänge zum subjektiven Glaubhaftigkeitsurteil ($R = .58$) als zum objektiven Wahrheitsstatus ($R = .43$) auf.

Dasselbe Stimulusmaterial wurde später von Reinhard et al. (2002) verwendet. Insgesamt 40 naive Beurteiler schätzten zunächst die Glaubhaftigkeit von jeweils zehn Aussagen anhand 10-stufiger Skalen ein. Danach formulierten sie freie Urteilsbegründungen, die später kategorisiert wurden. Von diesen Kategorien wurden die fünf häufigsten Gründe einer Brunswikschen Linsenmodellanalyse unterzogen. Die ökologischen Validitäten der Urteilsbegründungen waren gering ($R = .21$). Sie wiesen jedoch bedeutsame Zusammenhänge zu den subjektiven Glaubhaftigkeitsurteilen auf ($R = .52$). Entsprechend zeigte sich nur ein geringer Zusammenhang zwischen dem objektiven Wahrheitsstatus und der subjektiven Glaubhaftigkeitsbeurteilung ($r(198) = .15$).

Auch die ARJS-Merkmale wurden bereits einer Brunswikschen Linsenmodellanalyse unterzogen (Sporer, 1998). Insgesamt 71 Aussagen wurden von zwei Beurteilern anhand der ARJS bewertet. Zudem gaben beide intuitive

Glaubhaftigkeitsurteile auf 10-stufigen Skalen ab. Die Beurteiler stützten sich bei ihren Urteilen recht stark auf die 13 ARJS-Merkmale (R = .89). Für manche Merkmale ergaben sich allerdings nur geringe ökologische Validitäten. Daher fiel der Zusammenhang zwischen den ARJS-Merkmalen insgesamt und dem objektiven Wahrheitsstatus (R = .59) geringer aus. Es zeigte sich eine bedeutsame Korrelation von r = .35 (p < .001) zwischen dem objektiven Wahrheitsstatus und dem subjektiven Glaubhaftigkeitsurteil.

Cramer (2005; vgl. Sporer & Masip, 2007) ließ 32 Beurteiler jeweils 16 Aussagen zu moralischen oder gesetzlichen Vergehen bewerten. Die Burteiler wurden schriftlich über die Merkmale der ARJS-Kurzform informiert. Anhand der dazugehörigen 17 Merkmale führten sie eine Aussageanalyse durch und gaben anschließend 10-stufige Glaubhaftigkeitsurteile ab. Über 64 Aussagen hinweg war kein Zusammenhang zwischen dem objektiven Wahrheitsstatus und dem subjektiven Glaubhaftigkeitsurteil festzustellen (r(510) = .01). Zudem wurden die einzelnen Merkmale der ARJS-Kurzform einerseits mit dem objektiven, andererseits mit dem subjektiven Glaubhaftigkeitsurteil korreliert. Nur zwei Aussagemerkmale (r(510) = .22 für negative Äußerungen über sich selbst; r(510) =.14 für Eingeständnis von Erinnerungslücken) wiesen signifikante Zusammenhänge zum objektiven Wahrheitsstatus auf (-.10 < r(510) < .08 für alle übrigen Merkmale). Hingegen nutzten die Beurteiler sämtliche Merkmale (.13 < r(510) < .39), um ein Glaubhaftigkeitsurteil abzuleiten. Die geringe Urteilsgüte war demnach auf die unzureichenden ökologischen Validitäten der Merkmale zurückzuführen.

Untersuchungen anhand des Brunswikschen Linsenmodells machen deutlich, wie wichtig es ist zwischen der Validität von Aussagemerkmalen und der Güte der subjektiven Glaubhaftigkeitsurteile zu unterscheiden. Wiederholt gingen die untersuchten Aussagemerkmale stärker in das subjektive Glaubhaftigkeitsurteil mit ein, als es aufgrund ihrer ökologischen Validitäten sinnvoll erschien. Verbesserungen in der Urteilsgüte sind jedoch nur dann zu

erwarten, wenn die trainierten Merkmale auch für das beurteilungsrelevante Stimulusmaterial valide sind.

Ziele und Hypothesen

Die vorliegende Untersuchung überprüfte die Effekte einer kurzen Anleitung anhand der Kurzform der ARJS auf die Urteilsgüte, die Urteilsneigung und die subjektive Urteilssicherheit. Bisherige Studien untersuchten zwar häufig die Validität inhaltlicher Merkmale, jedoch nur selten die Richtigkeit der subjektiven Glaubhaftigkeitsurteile. Wenn sich Angaben hierzu finden, basieren diese oft nur auf den Bewertungen durch wenige Experten. Doch selbst bei erfahrenen und intensiv geschulten Beurteilern sind interindividuelle Unterschiede zu beachten. Beispielsweise schulte Sporer (1997a) zwei Beurteiler gleichermaßen intensiv in der Aussageanalyse anhand von 13 CBCA- und acht RÜ-Merkmalen. Dennoch fielen ihre subjektiven Glaubhaftigkeitsurteile unterschiedlich aus. Dies galt sowohl für die Urteilsgüte als auch für die Urteilsneigung. Daher wurden in der vorliegenden Untersuchung die Einschätzungen von insgesamt acht Beurteilerinnen analysiert (vgl. Studie 4). Um Trainingseffekte anhand eines Within-Subjects-Designs zu überprüfen, beurteilten sie eine Hälfte des Stimulusmaterials naiv, die andere nach einer kurzen ARJS-STV-G-Anleitung. Das Stimulusmaterial umfasste wahre und erfundene Aussagen zu bedeutsamen Lebensereignissen. Diese wurden entweder unvorbereitet oder nach einer kurzen Vorbereitungszeit formuliert.

Gemäß der Befunde zur Urteilsgüte von Laien wurde erwartet, dass die naiven Beurteilerinnen eine geringfügig überzufällige Urteilsrichtigkeit erzielen (Hypothese 1a). Des Weiteren wurde postuliert, dass die naiven Beurteilerinnen einen Wahrheitsbias zeigen (Hypothese 1b) und ihre Urteilsgüte überschätzen (Hypothese 1c).

Die ARJS-STV-G (ARJS--Short Training Version--German) stellt eine Übersetzung der englischsprachigen und spanischsprachigen Kurzform der ARJS

dar (Sporer & Masip, 2007). Sie umfasst 17 empirisch fundierte Glaubhaftigkeitsmerkmale. Ihre Validität wurde für das vorliegende Stimulusmaterial bereits nachgewiesen (vgl. Studie 4). Daher ist anzunehmen, dass die Beurteilerinnen durch eine ARJS-STV-G-Anleitung ihre Fähigkeit verbessern, wahre und erfundene Aussagen zu unterscheiden (Hypothese 2a). Entsprechend wird eine Wechselwirkung zwischen der Anleitung und dem Wahrheitsstatus auf die subjektiven Glaubhaftigkeitsurteile erwartet. Werden die subjektiven Glaubhaftigkeitsurteile dichotomisiert, sollten sie nach der ARJS-STV-G-Anleitung häufiger mit dem objektiven Wahrheitsstatus korrespondieren als davor. So wird eine höhere Urteilsrichtigkeit für die ARJS-STV-G-geleiteten im Vergleich zu den naiven Beurteilungen postuliert.

Des Weiteren wird angenommen, dass die Urteilsneigung der Beurteilerinnen durch die Anleitung beeinflusst wird. Durch die ARJS-STV-G-Anleitung werden sie für Merkmale sensibilisiert, die für einen persönlichen Erlebnisbezug sprechen. Daher wird eine Verstärkung des Wahrheitsbias erwartet. Die Tendenz Aussagen als glaubhaft zu beurteilen sollte mit der Anleitung stärker ausgeprägt sein als ohne die Anleitung (Hypothese 2b). Demnach wird ein Haupteffekt der Anleitung auf die subjektiven Glaubhaftigkeitsurteile postuliert.

Für die naiven Beurteilerinnen ist aufgrund des bisherigen Forschungsstandes kein Zusammenhang zwischen der Urteilsgüte und der subjektiven Sicherheit zu erwarten (vgl. Aamodt & Custer, 2006; DePaulo et al., 1997). Durch die ARJS-STV-G-Anleitung haben sie jedoch konkrete Hilfestellungen für die Aussageanalyse und die Ableitung des subjektiven Glaubhaftigkeitsurteils erhalten. Daher wird eine bessere Korrespondenz zwischen der Urteilsrichtigkeit und der subjektiven Sicherheit erwartet (Hypothese 2c). Dies würde sich in einem signifikanten Zusammenhang zwischen der Urteilsrichtigkeit und der subjektiven Sicherheit für die ARJS-STV-G-angeleiteten Beurteilungen widerspiegeln.

Insgesamt wurde angenommen, dass die Glaubhaftigkeitseinschätzung durch eine Vorbereitung der Aussage erschwert wird. Dies sollte die Urteilsrichtigkeit der naiven und ARJS-STV-G geschulten Beurteiler gleichermaßen beeinträchtigen. Daher wurde ein Haupteffekt der Vorbereitung auf die dichotome Urteilsrichtigkeit postuliert (Hypothese 3).

Zusätzlich soll die Urteilsgüte der ARJS-STV-G-geschulten Beurteilerinnen mit der von Experten verglichen werden. In einer früheren Untersuchung (vgl. Studie 3) bewerteten vier Beurteilerinnen dasselbe Stimulusmaterial anhand der Langform der ARJS und gaben ebenfalls Glaubhaftigkeitsurteile ab. Diese Beurteilerinnen verfügten über fundiertes theoretisches Wissen und absolvierten vor der Datenerhebung eine intensive Schulung. Daher wurde angenommen, dass diese Experten eine höhere Urteilsgüte erzielen als die anhand der Kurzform trainierten Laien (Hypothese 4). Dies sollte sich anhand der prozentualen Urteilsrichtigkeiten und der Diskriminationsleistung A' nachweisen lassen.

Schließlich soll eine Brunswiksche Linsenmodellanalyse durchgeführt werden, um die Urteilsgüte und die Validität der Glaubhaftigkeitsmerkmale einander vergleichend gegenüberzustellen. Die ökologischen Validitäten wurden bereits in den Studien 3 und 4 ausführlich besprochen. Die vorliegende Untersuchung soll hingegen klären, inwieweit diese Merkmale in das subjektive Glaubhaftigkeitsurteil eingehen. Den Beurteilerinnen wurden konkrete Richtlinien für die Ableitung des Glaubhaftigkeitsurteils aufgrund der Merkmalsbewertungen vorgegeben. Daher wird eine starke Korrespondenz zwischen den Merkmalsbewertungen und dem subjektiven Glaubhaftigkeitsurteil erwartet (Hypothese 5).

Methode

In der vorliegenden Untersuchung wurden subjektive Glaubhaftigkeitsurteile analysiert, die im Rahmen der Studien 3 und 4 zur Validität der ARJS bzw. ARJS-STV-G erhoben wurden. Im Folgenden wird nur kurz auf relevante methodische

Aspekte eingegangen. Nähere Informationen zu den Glaubhaftigkeitsmerkmalen und der Durchführung der beiden Untersuchungen sind den Studien 3 und 4 zu entnehmen.

Stimulusmaterial

Das Stimulusmaterial umfasste transkribierte Aussagen zu persönlich bedeutsamen Ereignissen (vgl. Sporer & Burghardt, 2004; Studie 3; Studie 4). Die Themenwahl wurde den Stimuluspersonen weitestgehend selbst überlassen. Infolgedessen wurden sowohl positive als auch negative Lebensereignisse geschildert. Jede Stimulusperson gab zunächst einen freien Bericht ab und wiederholte ihre Aussage eine Woche später während eines Interviews. Unmittelbar vor den freien Berichten wurden der Wahrheitsstatus (erfunden versus wahr) und die Vorbereitungszeit (2 versus 15 Minuten) als Between-Subjects-Faktoren experimentell variiert.

Erhebung der Glaubhaftigkeitsurteile und der Urteilssicherheit

In zwei Studien wurden 176 freie Berichte und 176 Interviews durch acht bzw. vier unabhängige Beurteiler bewertet. Die subjektiven Glaubhaftigkeitsurteile wurden anhand 10-stufiger Skalen mit den Endpunkten 1 = erfunden/verfälscht und 10 = erlebt/wahr erhoben. Dabei indizieren Werte von 1-5 eine Beurteilung als nicht-glaubhaft, Werte von 6-10 als glaubhaft. Zudem gaben sämtliche Beurteilerinnen an, wie sicher sie sich ihrer Entscheidung waren. Die subjektive Sicherheit wurde über 5-stufige Beurteilungsskalen mit den Endpunkten 1 = sehr unsicher bis 5 = sehr sicher erfasst.

Durchführung der Trainingsstudie

Insgesamt acht Psychologiestudierende im Grundstudium nahmen als Beurteilerinnen an einer Trainingsstudie teil (vgl. Studie 4). Sie verfügten über keinerlei Vorkenntnisse der rechtspsychologischen Täuschungsliteratur. Zunächst beurteilten sie jeweils ein Viertel der freien Berichte und die Hälfte der Interviews naiv. Danach erhielten sie eine 2,5-stündige Anleitung zu den ARJS-STV-G.

Zunächst wurden die 17 ARJS-STV-G-Merkmale vorgestellt und ausführlich besprochen. Danach wurde anhand einer Sammlung von 62 Beispielsätzen die Zuordnung und Beurteilung der Merkmale eingeübt. Dadurch ließen sich Schwierigkeiten im Verständnis der einzelnen ARJS-STV-G-Merkmale und ihrer Abgrenzung voneinander klären. Schließlich wurde erläutert, wie aufgrund der Merkmalsbeurteilungen ein Gesamturteil hinsichtlich der Glaubhaftigkeit abzuleiten ist. Dazu werden die 17-ARJS-STV-G-Beurteilungen zunächst zu drei Gruppenurteilen zusammengefasst. Diese indizieren, ob die Merkmale der jeweiligen Gruppe eher sehr hoch (+), mittel (0) oder nicht hoch (-) ausgeprägt sind. Die drei Gruppenurteile sind wiederum mit unterschiedlichem Gewicht zu einem abschließenden Glaubhaftigkeitsurteil zu integrieren. Die drei ARJS-STV-G-Merkmale Klarheit und Lebendigkeit, Realismus und logische Struktur sowie persönliche Signifikanz bilden die Gruppe mit dem geringsten Gewicht. Doppelt zu gewichten ist das Gruppenurteil, das acht ARJS-STV-G-Merkmale zu Detailinformationen, Gedanken, Gedächtnisprozesse, sensorische Eindrücke, Gefühle und Interaktionen umfasst. Das Gruppenurteil für die übrigen sechs Merkmale wird dreifach gewichtet. Dabei handelt es sich um die Merkmale negative Äußerungen über sich selbst, spontane Korrekturen, Erinnerungslücken zugeben, Komplikationen, ungewöhnliche und überflüssige Details. Schließlich führten alle Beurteilerinnen für eine Übungsaussage eine ARJS-STV-G-Analyse durch. Die Integration der ARJS-STV-G-Beurteilungen zu den drei Gruppenurteilen und zu einem anschließenden Glaubhaftigkeitsurteil wurde für diese Übungsaussage besprochen.

Expertenurteile

In einer unabhängigen Studie beurteilten vier andere Beurteilerinnen dasselbe Stimulusmaterial (vgl. Studie 3). Jede bearbeitete jeweils die Hälfte der freien Berichte und sämtliche Interviews. Die Beurteilerinnen studierten Psychologie im Hauptstudium und verfügten bereits im Vorfeld der Untersuchung

über umfangreiches rechtspsychologisches Literaturwissen. Zudem beschäftigten sie sich in einem 50-stündigen Training intensiv mit der ARJS-Aussageanalyse. Gemäß des ARJS-Manuals wurden sie auf Beurteilungsfehler und die historische Entwicklung der ARJS hingewiesen (Sporer, 1996/1998/2004). Dabei wurden relevante Theorien und empirische Forschungsbefunde ausführlich dargestellt. Zudem übten sie die ARJS-Aussageanalyse anhand von 14 Transkripten und klärten anschließend Beurteilungsdiskrepanzen. Aufgrund ihrer theoretischen Kenntnisse und der intensiven Schulung werden sie als ARJS-Experten aufgefasst.

Die Langform der ARJS bietet ebenso wie die Kurzform konkrete Richtlinien zur Ableitung eines abschließenden Glaubhaftigkeitsurteils. Zunächst werden die 52 ARJS-Merkmale unabhängig voneinander bewertet. Diese Itembeurteilungen werden dann zu 15 Skalenurteilen zusammengefasst, aus denen wiederum drei Skalenblöcke gebildet werden. Dabei wird jeweils eingeschätzt, ob die zugrundeliegenden Merkmale gering (-), mittel (0) oder stark (+) ausgeprägt sind. Erneut gehen die drei Skalenblöcke mit unterschiedlicher Gewichtung in das Gesamturteil ein. Der Skalenblock mit dem geringsten Gewicht umfasst die Skalen Realismus und logische Struktur sowie Klarheit und Lebendigkeit. Im Gegensatz zur Kurzform der ARJS gehen Beurteilungen zur persönlichen Signifikanz der Ereignisse nicht in das Glaubhaftigkeitsurteil mit ein. Doppelt zu gewichten ist der Block, der sich auf Detailinformationen, Sinneseindrücke, Gefühle, Gedanken, Memorieren und Gedächtnis sowie Interaktionen bezieht. Die Skalen Komplikationen, ungewöhnliche Details, Fehler und sozial Unerwünschtes bilden den dreifach zu gewichtenden Block.

Ergebnisse

Zunächst wurden die Effekte der experimentellen Variationen auf die Glaubhaftigkeitsurteile, Urteilsgüte und -neigung varianzanalytisch überprüft. Der Wahrheitsstatus (erfunden versus wahr) und die Vorbereitung (2 versus 15

Minuten) stellten die beiden Between-Subjects-Faktoren der nachfolgenden 2 x 2 (x 2) ANOVAs dar, die Anleitung anhand der ARJS-STV-G (ohne versus mit Anleitung) den Within-Subjects-Faktor.

Als Effektstärkemaß wurde r berechnet. Dabei wurden unterschiedliche Formeln für Between- und Within-Subjects-Faktoren verwendet (vgl. Dunlap, Cortina, Vaslow & Burke, 1996; Lipsey & Wilson, 2001). Die Effektstärken für die Haupteffekte der Between-Subjects-Faktoren Wahrheitsstatus und Vorbereitung wurden anhand der F-Werte abgeleitet (r=SQRT((F/(F+df))), nach Mullen, 1989, p. 44). Für Effekte des Within-Subjects-Faktors wurde zunächst Cohen's d (d=(M_1-M_2)/SD_{pooled}, nach Cohen, 1988, p. 21) ermittelt und dann in r transformiert (r=d/(SQRT(d^2+4)), nach Mullen, 1989, p. 46). Bei Wechselwirkungen wurde f als Effektstärkemaß verwendet. Hypothesenkonforme Effekte sind durch positive Effektstärken gekennzeichnet, erwartungskonträre durch negative.

Die Analyseeinheit bildeten die N = 176 freien Berichte und N = 176 Interviews. Für jeden freien Bericht lagen jeweils zwei naive und ARJS-STV-G-geleitete Beurteilungen vor. Hingegen wurde jedes Interview von vier Raterinnen naiv und mit der ARJS-STV-G-Anleitung bearbeitet. Für die nachfolgenden Analysen wurden die Urteile der Raterinnen gemittelt, die dieselben Aussagen zum selben Zeitpunkt beurteilten. Entsprechend wurden für die freien Berichte die Beurteilungen von zwei, für die Interviews von vier Raterinnen zusammengefasst.

Die Ergebnisse werden im Folgenden für die freien Berichte und die Interviews getrennt berichtet. Laut Hypothesen postulierte Wechselwirkungen mit dem Within-Subjects-Faktor Anleitung werden zudem graphisch veranschaulicht.

Subjektive Glaubhaftigkeitsurteile

Zunächst wurden die anhand 10-stufiger Skalen erhobenen subjektiven Glaubhaftigkeitsurteile untersucht. Die Wechselwirkungen zwischen der ARJS-STV-G-Anleitung einerseits und dem Wahrheitsstatus bzw. der Vorbereitung andererseits waren dabei von besonderem Interesse. Daher wurden auch bei

nicht signifikanten varianzanalytischen Befunden die einfachen Haupteffekte überprüft.

Freie Berichte

Für die freien Berichte ergaben sich signifikante Haupteffekte des Wahrheitsstatus und der Vorbereitung auf das subjektive Glaubhaftigkeitsurteil. Wahre Aussagen (M = 6.13) wurden im Vergleich zu erfundenen (M = 5.69) als glaubhafter eingeschätzt, F(1,172) = 4.88, p = .029, r = .17. Zudem erzielten vorbereitete Aussagen höhere Glaubhaftigkeitsbeurteilungen (M = 6.13) als unvorbereitete (M = 5.69), F(1,172) = 5.00, p = .027, r = .17. Eine Wechselwirkung zwischen den beiden Between-Subjects-Faktoren war nicht festzustellen, F(1,172) = 0.04, p = .841, f = .02.

Die naiven (M = 5.79) und ARJS-STV-G-geleiteten Glaubhaftigkeitsurteile (M = 6.03) waren vergleichbar, F(1,172) = 2.35, p = .127, r = .07. Auch die Wechselwirkung zwischen der Anleitung und dem Wahrheitsstatus verfehlte statistische Signifikanz, F(1,172) = 3.18, p = .076, f = .14. Dennoch fielen die einfachen Haupteffekte des Wahrheitsstatus erwartungsgemäß aus (vgl. Abbildung 5.2). Die naiven Glaubhaftigkeitsurteile unterschieden sich nicht signifikant bei wahren (M = 5.86) und erfundenen Aussagen (M = 5.71), F(1,174) = 0.33, p = .569, r = .04. Mit der ARJS-STV-G-Anleitung ergaben sich hingegen hypothesenkonform höhere Bewertungen für wahre (M = 6.39) als für erfundene Aussagen (M = 5.67), F(1,174) = 8.76, p = .004, r = .22. Auch die Befunde zu den einfachen Haupteffekten der Anleitung waren erwartungsgemäß. Für erfundene Aussagen zeigten sich keine Unterschiede zwischen den naiven (M = 5.71) und ARJS-STV-geleiteten Beurteilungen (M = 5.67), t(87) = 0.21, p = .833, r = .01. Jedoch erzielten wahre Aussagen mit der ARJS-STV-G-Anleitung (M = 6.39) höhere Glaubhaftigkeitsbeurteilungen als ohne diese (M = 5.86), t(87) = 2.02, p = .047, r = .15.

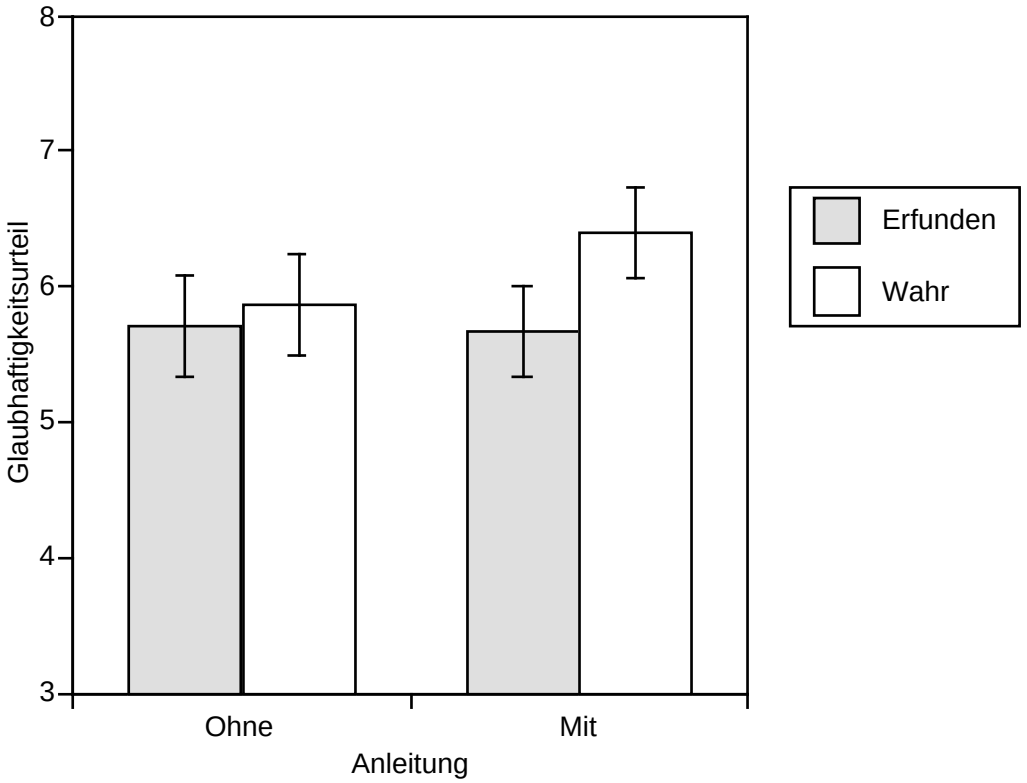


Abbildung 5.2: Mittelwerte (und 95% CIs) der Glaubhaftigkeitsurteile für erfundene und wahre Aussagen als Funktion der Anleitung (N = 176 freie Berichte).

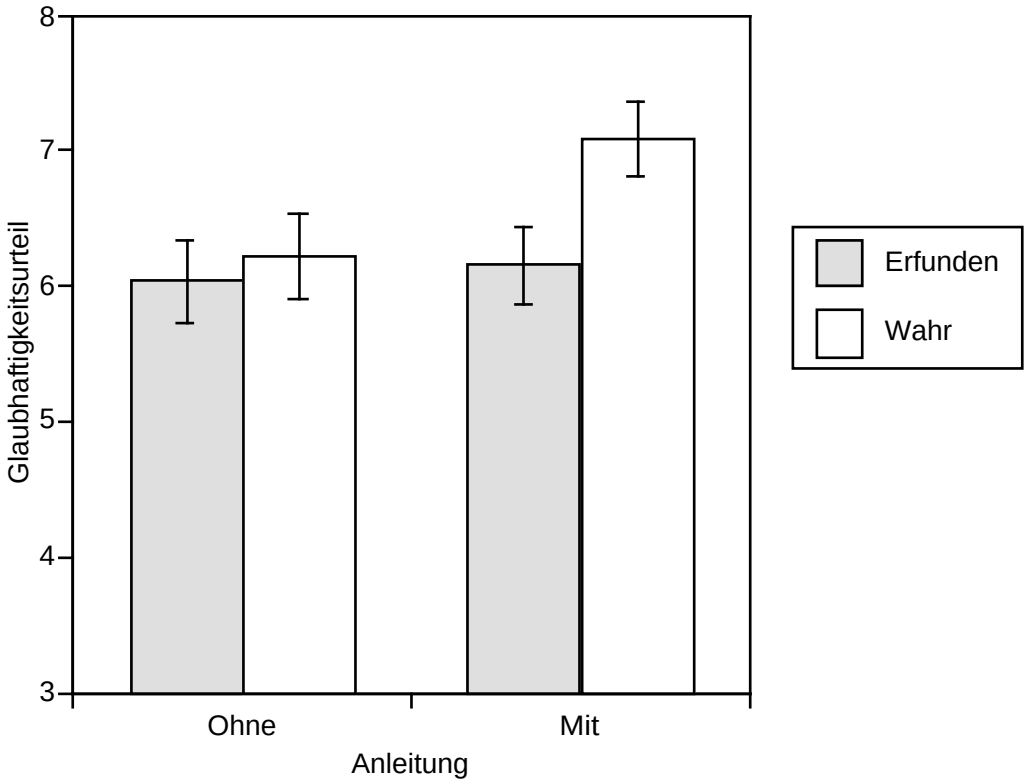


Abbildung 5.3. Mittelwerte (und 95% CIs) der Glaubhaftigkeitsurteile für erfundene und wahre Aussagen als Funktion der Anleitung (N =176 Interviews).

Des Weiteren zeigte sich eine signifikante Wechselwirkung zwischen der Anleitung der Beurteiler und der Vorbereitung der Aussagen, F(1,172) = 5.70, p = .018, f = .18. Die Mittelwerte und Konfidenzintervalle sind in Abbildung 5.4 dargestellt. Die Analyse der einfachen Haupteffekte der Anleitung ergab keine signifikanten Unterschiede für unvorbereitete Aussagen (ohne Anleitung: M = 5.76, mit Anleitung: M = 5.62), t(87) = -0.64, p = .521, r = .04. Für vorbereitete Aussagen wurden jedoch mit der ARJS-STVG-Anleitung höhere Glaubhaftigkeitsurteile (M = 6.44) vergeben als ohne Anleitung (M = 5.82), t(87) = 2.61, p = .011, r = .17.

Es lag keine Wechselwirkung zwischen den drei Faktoren Anleitung, Wahrheitsstatus und Vorbereitung vor, F(1,172) = .046, p = .831, f = .02.

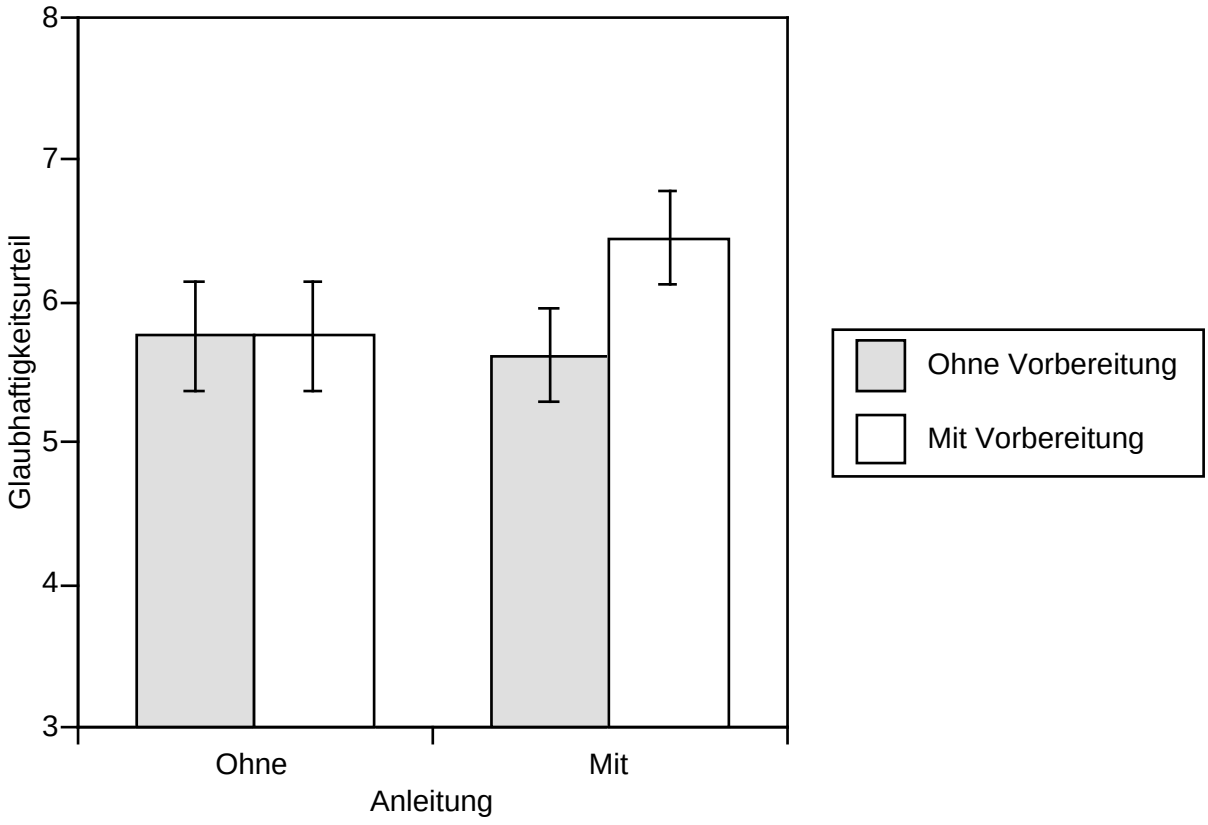


Abbildung 5.4: Mittelwerte (und 95% CIs) der Glaubhaftigkeitsurteile für Aussagen mit vs. ohne Vorbereitung als Funktion der Anleitung (N =176 freie Berichte)

Interviews

Für die Interviews ergab sich ebenfalls ein signifikanter Haupteffekt des Wahrheitsstatus. Wahre Aussagen (M = 6.65) wurden im Vergleich zu erfundenen (M = 6.10) als glaubhafter beurteilt, F(1,172) = 10.25, p = .002, r = .24. Es gab jedoch keine Unterschiede in der Einschätzung vorbereiteter (M = 6.36) und unvorbereiteter Aussagen (M = 6.39), F(1,172) = 0.02, p = .903, r = -.01. Auch war keine Wechselwirkung der beiden Between-Subjects-Faktoren festzustellen, F(1,172) = 0.02, p = .890, f = .01.

Im Gegensatz zu den freien Berichten zeigte sich für die Interviews der erwartete Effekt der Anleitung auf die Glaubhaftigkeitsbeurteilung. Mit der Anleitung zu den ARJS-STV-G-Merkmalen (M = 6.62) wurden höhere Glaubhaftigkeitsbeurteilungen vergeben als ohne diese (M = 6.13), F(1,172) = 16.45, p < .001, r = .17. Zudem interagierte die Anleitung mit dem Wahrheitsstatus, F(1,172) = 10.08, p = .002, f = .24. Die Mittelwerte und Konfidenzintervalle der subjektiven Glaubhaftigkeitsurteile sind in Abbildung 5.3 dargestellt. Die naiven Glaubhaftigkeitsurteile waren bei wahren (M = 6.22) und erfundenen Aussagen (M = 6.04) vergleichbar, F(1,174) = 0.66, p = .418, r = .06. Mit der ARJS-STV-G-Anleitung erzielten wahre Aussagen (M = 7.09) jedoch deutlich höhere Bewertungen als erfundene (M = 6.15), F(1,174) = 21.84, p < .001, r = .33. Es zeigten sich keine Unterschiede in der naiven und ARJS-STV-G-geleiteten Beurteilung von erfundenen Aussagen (M = 6.04 und M = 6.15, respektive), t(87) = -0.64, p = .522, r = -.03. Für wahre Aussagen ergab sich jedoch ein erwartungsgemäßer einfacher Haupteffekt der Anleitung (ohne Anleitung: M = 6.22, mit Anleitung: M = 7.09), t(87) = 4.92, p < .001, r = .32.

Es war keine Wechselwirkung zwischen der Anleitung und der Vorbereitung festzustellen, F(1,172) = 1.17, p = .281, f = .08. Die beiden einfachen Haupteffekte der Anleitung waren jedoch signifikant. So wurden unvorbereitete Aussagen mit der ARJS-STV-G-Anleitung (M = 6.56) als glaubhafter eingeschätzt als ohne die Anleitung (M = 6.21), t(87) = -2.25, p = .027, r = -.13. Auch vorbereitete Aussagen

erzielten höhere Glaubhaftigkeitsbeurteilungen bei der ARJS-STV-G-geleiteten (M = 6.67) im Vergleich zur naiven Beurteilung (M = 6.06), t(87) = 3.25, p = .002, r = .20.

Erneut war keine Wechselwirkung zwischen allen drei Faktoren nachweisbar, F(1,172) = 2.43, p = .121, f = .12.

Richtigkeit der Entscheidungen

Um die Entscheidungsrichtigkeit zu bestimmen, wurden die intervallskalierten Glaubhaftigkeitsurteile dichotomisiert. Die 10-stufige Beurteilungsskala ermöglicht es für jede einzelne Raterin zwischen einer Beurteilung als nicht-glaubhaft (Werte von 1-5) und als glaubhaft (Werte von 6-10) zu differenzieren. Durch die Zusammenfassung der Urteile mehrerer Raterinnen resultierten jedoch nicht-entscheidbare Fälle. So ergab sich für n = 16 (9.1%) der naiv und n = 13 (7.4%) der ARJS-STV-G-geleitet beurteilten freien Berichte ein durchschnittliches Glaubhaftigkeitsurteil von 5.5. Für die Interviews erzielten n = 8 (4.5%) Aussagen vor und n = 5 (2.8%) Aussagen nach der ARJS-STV-G-Anleitung eine Beurteilung von 5.5. Diese Fälle wurden den als glaubhaft beurteilten Aussagen zugeordnet. Glaubhaftigkeitsurteile zwischen 1 und 5.25 repräsentieren demnach eine Beurteilung als nicht-glaubhaft. Hingegen indizieren Werte zwischen 5.5 und 10, dass die entsprechenden Aussagen als glaubhaft eingeschätzt wurden. Die so abgeleiteten dichotomen Entscheidungen für bzw. gegen die Glaubhaftigkeit wurden mit dem tatsächlichen Wahrheitsstatus der Aussagen abgeglichen. Daraus ergab sich eine neue dichotome Variable, welche die Entscheidungsrichtigkeit widerspiegelt. Falschen Entscheidungen wurde der Wert 0, richtigen Entscheidungen der Wert 1 zugeordnet. Mit 100 multipliziert ergeben diese die prozentualen Urteilsrichtigkeiten. Dies soll die Vergleichbarkeit mit relevanten Studien erleichtern.

Freie Berichte

Der Wahrheitsstatus hatte einen signifikanten Effekt auf die Richtigkeit der Entscheidung. Wahre Aussagen (69.9%) wurden häufiger richtig beurteilt als

erfundene (37.5%), F(1,172) = 40.41, p < .001, r = .44. Allerdings war kein Effekt der Vorbereitung auf die Entscheidungsrichtigkeit festzustellen (ohne Vorbereitung: 54.5%, mit Vorbereitung: 52.8%), F(1,172) = 0.11, p = .738, r = .03. Die Wechselwirkung zwischen den beiden Between-Subjects-Faktoren war erneut nicht signifikant, F(1,172) = 2.10, p = .149, f = .11.

Die Anleitung hatte keinen signifikanten Effekt auf die Richtigkeit der Entscheidungen (ohne Anleitung: 50.6%, mit Anleitung: 56.8%), F(1,172) = 1.63, p = .203, r = .06. Auch zeigte sich keine Interaktion zwischen der Anleitung und dem Wahrheitsstatus (vgl. Abbildung 5.5), F(1,172) = 1.63, p = .203, f = .10. Erfundene Aussagen wurden mit der ARJS-STV-G-Anleitung (37.5%) ebenso selten richtig eingeschätzt, wie ohne diese (37.5%), t(87) = 0.00, p = 1.000, r = .00. Bei wahren Aussagen war der Unterschied zwischen der Entscheidungsrichtigkeit für die naiven (63.6%) und ARJS-STV-G-geleiteten Beurteilungen (76.1%) marginal signifikant, t(87) = 1.83, p = .070, r = .13.

Schließlich war keine Wechselwirkung zwischen der Anleitung und der Vorbereitung nachweisbar, F(1,172) = 0.12, p = .728, f = .03. Unvorbereitete Aussagen wurden naiv (52.3%) und mit der ARJS-STV-G-Anleitung (56.8%) gleichermaßen häufig richtig eingeschätzt, t(87) = 0.60, p = .550, r = .05. Doch auch für vorbereitete Aussagen waren keine anleitungsbedingten Unterschiede in der Entscheidungsrichtigkeit festzustellen (ohne Anleitung: 48.9%, mit Anleitung: 56.8%), t(87) = 1.22, p = .225, r = .08.

Es ergab sich jedoch eine Interaktion zwischen allen drei Faktoren, F(1,172) = 7.12, p = .008, f = .20. Die Abbildungen 5.7a und 5.7b zeigen die Urteilsrichtigkeit ohne und mit der ARJS-STV-G-Anleitung in Abhängigkeit von der Vorbereitung und dem Wahrheitsstatus der Aussagen. Bei den naiven Beurteilungen wurden erfundene Aussagen unabhängig von der Vorbereitung richtig eingeschätzt (unvorbereitete: 36.4%, vorbereitete: 38.6%), F(1,172) = 0.05, p = .827, r = .02. Mit der ARJS-STV-G-Anleitung zeigte sich jedoch eine geringere Entscheidungsrichtigkeit für vorbereitete (27.7%) im Vergleich zu unvorbereiteten

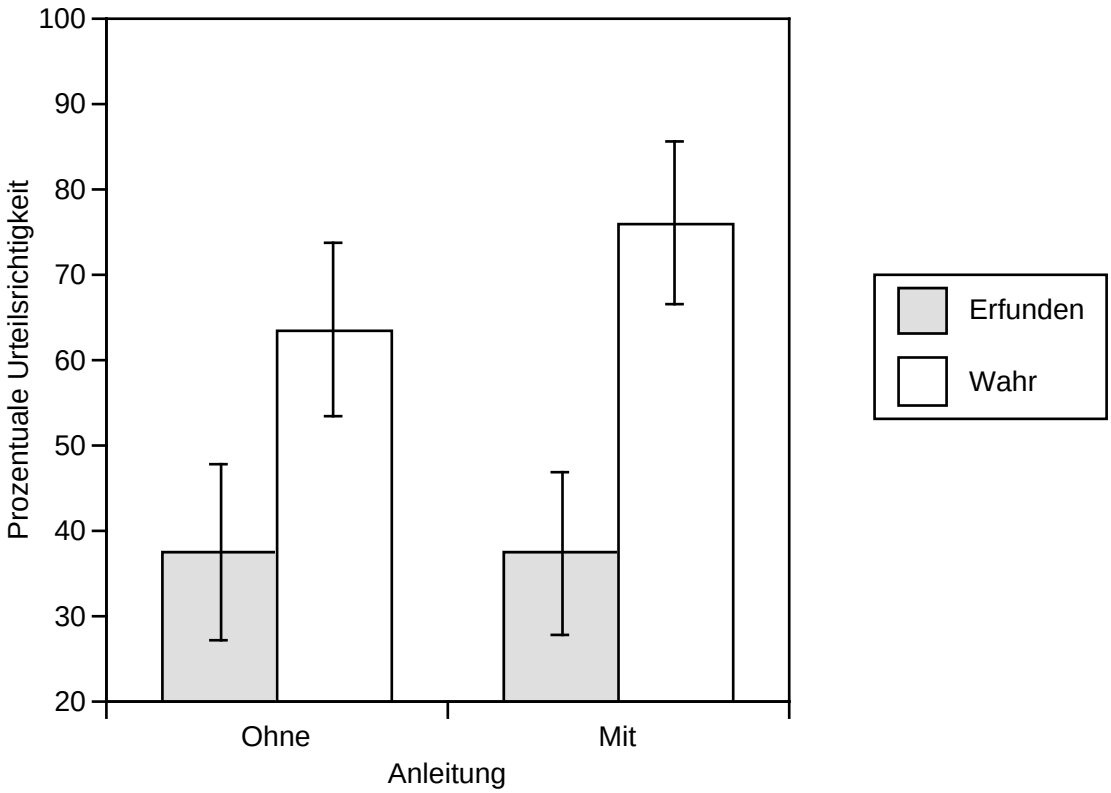


Abbildung 5.5: Mittelwerte (und 95% CIs) der prozentualen Urteilsrichtigkeit für erfundene und wahre Aussagen als Funktion der Anleitung (N =176 freie Berichte).

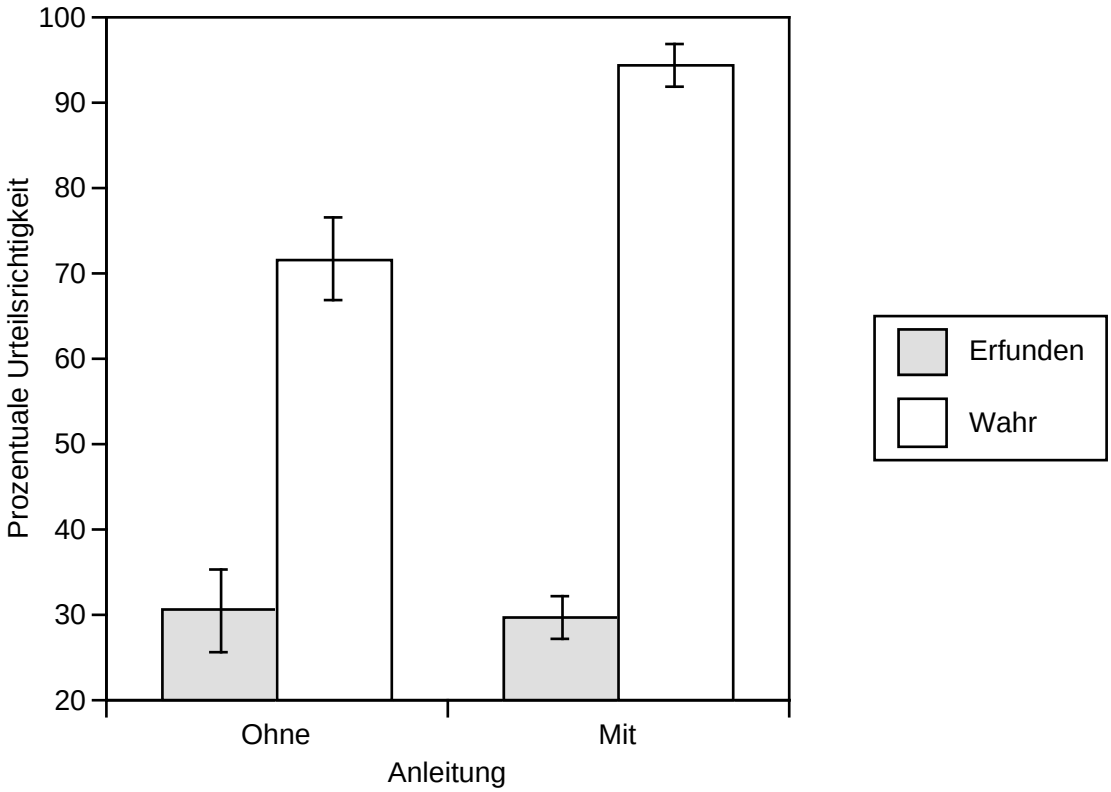


Abbildung 5.6: Mittelwerte (und 95% CIs) der prozentualen Urteilsrichtigkeit für erfundene und wahre Aussagen als Funktion der Anleitung (N =176 Interviews).

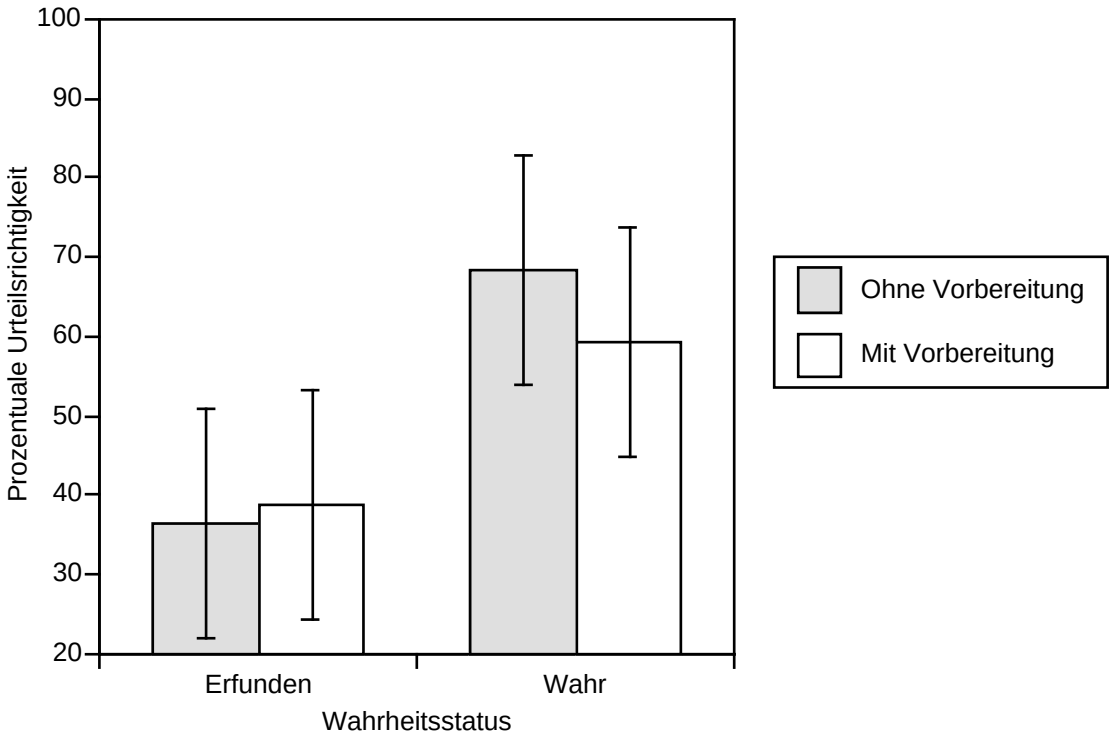


Abbildung 5.7a: Mittelwerte (und 95% CIs) der prozentualen Urteilsrichtigkeit als Funktion des Wahrheitsstatus und der Vorbereitungszeit ohne die ARJS-STV-G-Anleitung (N =176 freie Berichte).

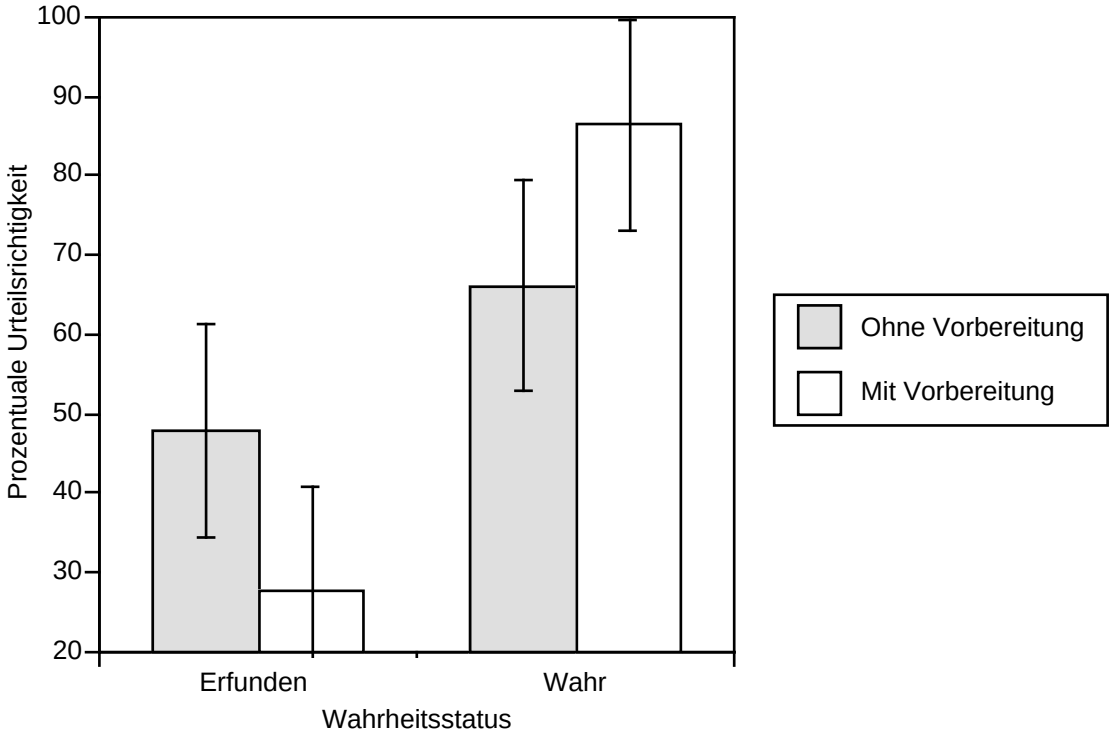


Abbildung 5.7b: Mittelwerte (und 95% CIs) der prozentualen Urteilsrichtigkeit als Funktion des Wahrheitsstatus und der Vorbereitungszeit mit der ARJS-STV-G-Anleitung (N =176 freie Berichte).

Lügen (47.7%), F(1,172) = 4.55, p = .034, r = .16. Auch bei wahren Aussagen waren für die naiven Beurteilungen kaum vorbereitungsbedingte Unterschiede in der Richtigkeit der Entscheidung festzustellen (unvorbereitete: 68.2%, vorbereitete: 59.1%), F(1,172) = 0.77, p = .383, r = .07. Mit der Anleitung wurden vorbereitete wahre Aussagen jedoch (86.4%) häufiger richtig beurteilt als unvorbereitete (65.9%), F(1,172) = 4.55, p = .034, r = .16.

Interviews

Für die Interviews war ebenfalls ein signifikanter Haupteffekt des Wahrheitsstatus auf die Entscheidungsrichtigkeit nachweisbar. Wahre Aussagen (83.0%) wurden häufiger richtig beurteilt als erfundene (30.1%), F(1,172) = 113.56, p < .001, r = .63. Die Urteilsrichtigkeit für vorbereitete (55.1%) und unvorbereitete Aussagen (58.0%) war vergleichbar, F(1,172) = 0.33, p = .567, r = -.04. Zudem zeigte sich keine Wechselwirkung zwischen den beiden Between-Subjects-Faktoren, F(1,172) = 0.64, p = .424, f = .06.

Im Gegensatz zu den freien Berichten ergab sich für die Interviews ein Haupteffekt der Anleitung auf die Richtigkeit der Entscheidungen. Ohne Anleitung wurden weniger richtige Entscheidungen getroffen (51.1%), als mit der Anleitung zu den Glaubhaftigkeitsmerkmalen (61.9%), F(1,172) = 7.99, p = .005, r = .11. Zudem war eine signifikante Interaktion zwischen der Anleitung und dem Wahrheitsstatus nachweisbar, F(1,172) = 9.76, p = .002, f = .24. Abbildung 5.6 ist zu entnehmen, dass erfundene Aussagen ohne (30.7%) und mit der ARJS-STV-G-Anleitung (29.5%) gleichermaßen selten richtig eingeschätzt wurden, t(87) = -0.21, p = .836, r = -.01. Hingegen wurden wahre Aussagen nach der ARJS-STV-G-Anleitung (94.3%) häufiger richtig beurteilt als ohne die Anleitung (71.6%), t(87) = -4.29, p < .001, r = .30.

Eine Wechselwirkung zwischen der Anleitung und der Vorbereitung war hingegen nicht zu erkennen, F(1,172) = 0.55, p = .458, f = .05. Dennoch fielen die einfachen Haupteffekte der Anleitung erwartungsgemäß aus. Die Entscheidungsrichtigkeit für unvorbereitete Aussagen war bei naiver Beurteilung

(51.1%) geringer als mit der ARJS-STV-G-Anleitung (64.8%), t(87) = 2.78, p = .007, r = .14. Für vorbereitete Aussagen unterschied sich die Entscheidungsrichtigkeit der Beurteiler ohne (51.1%) und mit der Anleitung (59.1%) hingegen nicht signifikant, t(87) = 1.31, p = .195, r = .08.

Für die Interviews war keine Wechselwirkung zwischen allen drei Faktoren festzustellen, F(1,172) = 0.55, p = .458, f = .05.

Quantifizierung der Urteilsgüte und -neigung

Die Befunde zur prozentualen Urteilsrichtigkeit der naiven und ARJS-STV-G-geleiteten Beurteilerinnen sind in Tabelle 5.5 zusammengefasst. Dabei wurden auch die prozentualen Häufigkeiten der als glaubhaft beurteilten wahren Aussagen (Treffer) und der als glaubhaft beurteilten erfundenen Aussagen (falsche Alarme) aufgeführt. Um zu überprüfen, ob die Werte als überzufällig zu interpretieren waren, wurden one-sample t-Tests durchgeführt. Als Testwert diente dabei die jeweils zufällig zu erwartende Urteilsrichtigkeit. Für die Urteilsrichtigkeit insgesamt war eine zufällige Entscheidungsrichtigkeit von 50% zu erwarten. Die zufällig zu erwartende Trefferrate entsprach dem Prozentsatz der als glaubhaft beurteilten Aussagen. Die zufällig zu erwartende Rate der korrekten Zurückweisungen wiederum entsprach dem Prozentsatz der als nicht-glaubhaft eingeschätzten Aussagen.

Tabelle 5.5
Prozentuale Urteilsrichtigkeit für wahre und erfundene Aussagen sowie Urteilsgüte A' und Urteilsneigung B" der naiven, ARJS-STV-G-geleiteten und Expertenbeurteilung

	Freie Berichte	Interviews
Naive Beurteilung	N = 176 (n = 160)	N = 176 (n = 168)
Beurteilungen als glaubhaft (%)[a]	**63.1 (59.4)**	**70.4 (69.1)**
Urteilsrichtigkeit insgesamt (%)[a]	50.6 (51.9)	51.1 (51.8)
Treffer (%)[b]	63.6 (61.0)	71.6 (70.6)
Korrekte Zurückweisungen (%)[b]	37.5 (42.3)	30.7 (32.5)
Urteilsgüte A'	.51 (.53)	.53 (.54)
Urteilsneigung B"	-.49 (-.36)	-.70 (-.67)
ARJS-STV-G-Beurteilung	N = 176 (n = 163)	N = 176 (n = 171)
Beurteilungen als glaubhaft (%)[a]	**69.3 (66.9)**	**82.4 (81.9)**
Urteilsrichtigkeit insgesamt (%)[a]	56.8 **(58.9)**	**61.9 (63.2)**
Treffer (%)[b]	76.1 (75.0)	**94.3 (94.3)**
Korrekte Zurückweisungen (%)[b]	37.5 (41.8)	**29.5 (31.0)**
Urteilsgüte A'	.64 (.66)	.77 (.77)
Urteilsneigung B"	-.68 (-.61)	-.95 (-.95)
Expertenbeurteilung	N = 176 (n = 163)	N = 176 (n = 166)
Beurteilungen als glaubhaft (%)[a]	**57.4** (54.0)	**68.1 (66.3)**
Urteilsrichtigkeit insgesamt (%)[a]	**64.2 (65.0)**	**67.0 (68.1)**
Treffer (%)[b]	**71.6 (69.1)**	**85.2 (84.3)**
Korrekte Zurückweisungen (%)[b]	**56.8 (61.0)**	**48.9 (51.8)**
Urteilsgüte A'	.72 (.73)	.77 (.78)
Urteilsneigung B"	-.31 (-.18)	-.72 (-.67)

Anm. Die Angaben basieren für die freien Berichte auf der mittleren Beurteilung durch zwei Raterinnen, für die Interviews durch vier Raterinnen.
[a] = fett gedruckte Werte unterscheiden sich signifikant von 50%;
[b] = fett gedruckte Werte unterscheiden sich signifikant von der aufgrund der Urteilsneigung zu erwartenden zufälligen Urteilsrichtigkeit. Werte in Klammern entsprechen den Befunden unter Ausschluss unentschiedener Urteile (mittleres Glaubhaftigkeitsurteil von 5.5).

Zudem wurden signalentdeckungstheoretische Kennwerte abgeleitet, um die Urteilsgüte und –neigung getrennt voneinander zu quantifizieren. Dazu wurden die Formeln von Rae (1976) und Donaldson (1992) verwendet. Die vorliegenden Untersuchungsbefunde ergänzend wurden auch die Befunde einer Untersuchung dokumentiert, in denen Experten dieselben Aussagen anhand der Langform der ARJS beurteilten (vgl. Studie 3).

Schließlich wurden sämtliche Analysen auch unter Ausschluss der Aussagen, die ein mittleres Glaubhaftigkeitsurteil von 5.5 erzielten durchgeführt. Die Ergebnisse sind ebenfalls in Tabelle 5.5 aufgeführt und werden später bei der Diskussion der Untersuchungsbefunde aufgegriffen.

Freie Berichte

Auch die signalentdeckungstheoretischen Analysen zeigten, dass die naiven Raterinnen keine überzeugende Urteilsfähigkeit aufwiesen (A' = .51) und ein liberales Antwortkriterium wählten (B" = -.49). Nach der ARJS-STV-G-Anleitung zeigte sich eine deutlich bessere Urteilsfähigkeit von A' = .76. Zudem war erneut ein ausgeprägter Wahrheitsbias feststellbar (B" = -.68).

Interviews

Für die Interviews zeigte sich ebenfalls nur eine geringe Urteilsgüte der naiven Beurteiler (A' = .53). Nach der ARJS-STV-G-Anleitung ergab sich jedoch eine stark verbesserte Urteilsgüte von A' = .77. Jedoch unterlagen die Raterinnen erneut sowohl bei den naiven (B" = -.70) als auch bei den ARJS-STV-G-geleiteten Beurteilungen einem deutlichen Wahrheitsbias (B" = -.95).

Subjektive Sicherheit

Schließlich wurde die subjektive Sicherheit der Beurteilerinnen analysiert, die anhand 5-stufiger Skalen erhoben worden war. Zunächst wurden die Zusammenhänge zur Entscheidungsrichtigkeit korrelativ überprüft. Zudem wurden

die Zusammenhänge zu dem dichotomisierten Glaubhaftigkeitsurteil der Raterinnen ermittelt.

Freie Berichte

Für die freien Berichte ergab sich eine mittlere subjektive Sicherheit (M = 3.34, SD = 0.61). Dies galt sowohl für die naiven (M = 3.35, SD = 0.69), als auch für die ARJS-STV-G-geleiteten Beurteilungen (M = 3.34, SD = 0.98), t(175) = 0.13, p = .897. Bei den naiven Beurteilungen ergab sich kein Zusammenhang zwischen der subjektiven Sicherheit und der Entscheidungsrichtigkeit, r(174) = -.06, p = .399. Mit der Anleitung zu den ARJS-STV-G-Merkmalen war jedoch ein positiver Zusammenhang von r(174) = .22, p = .004, festzustellen. Entscheidungen für die Glaubhaftigkeit von Aussagen gingen mit einer höheren subjektiven Sicherheit einher. Dieser Zusammenhang war sowohl bei den naiven Beurteilungen nachweisbar (r(174) = .19, p = .012) als auch bei den ARJS-STV-G-geleiteten (r(174) = .20, p = .007).

Interviews

Die Beurteiler waren sich auch bei den Interviews ihrer Entscheidungen sicher (M = 3.54, SD = 0.47). Mit der ARJS-STV-G-Anleitung (M = 3.64, SD = 0.72) wurde eine höhere Urteilssicherheit berichtet als bei den naiven Beurteilungen (M = 3.44, SD = 0.53), t(175)= 3.07, p = .003. Für die naiven Beurteilungen korrelierte die subjektive Sicherheit weder mit der Richtigkeit der Entscheidung (r(174) = .04, p = .598) noch mit dem Glaubhaftigkeitsurteil (r(174) = .09, p = .228). Auch mit der ARJS-STV-G-Anleitung war kein signifikanter Zusammenhang zwischen der Sicherheit und Richtigkeit der Entscheidungen festzustellen, r(174) = .13, p = .089. Die Beurteiler stuften ihre subjektive Sicherheit jedoch höher ein, wenn sie Aussagen als glaubhaft bewerteten, r(174) = .26, p = .001.

Brunswiksche Linsenmodellanalyse

Getrennt für die freien Berichte und die Interviews wurden Brunswiksche Linsenmodellanalysen durchgeführt. Als Prädiktoren dienten zum einen die 17

ARJS-STV-G-Merkmale. Die Beurteilerinnen integrierten deren Bewertungen zudem zu drei Kategorien, die mit unterschiedlicher Gewichtung in das abschließende Glaubhaftigkeitsurteil eingehen sollten. Diese integrativen Urteile wurden als Prädiktoren einer weiteren Brunswikschen Linsenmodellanalyse unterzogen. Die Beurteilung der Prädiktoren basierte für die freien Berichte auf der mittleren Einschätzung durch zwei, für die Interviews durch vier Raterinnen.

Für eine Brunswiksche Linsenmodellanalyse sind zwei multiple Regressionsanalysen durchzuführen. Für die erste Regressionsanalyse wurde der objektive Wahrheitsstatus als Kriterium verwendet. Die Befunde verweisen auf die Validität der Glaubhaftigkeitsmerkmale. Für die zweite Regressionsanalyse wurde das dichotomisierte subjektive Glaubhaftigkeitsurteil als Kriterium herangezogen. Dadurch wird der Zusammenhang zwischen den ARJS-STV-G-Beurteilungen und der Glaubhaftigkeitsattribution überprüft.

Anhand der Regressionsgewichte lässt sich abschätzen, inwieweit jeder Prädiktor eigenständig zur Varianzaufklärung des entsprechenden Kriteriums beiträgt. Der Einfluss der übrigen Prädiktoren wird dazu aus jedem Prädiktor herauspartialisiert. Dieses Vorgehen beseitigt Redundanzen und erlaubt es die relative Wichtigkeit einzelner Prädiktoren abzuschätzen. Die Interpretierbarkeit der Regressionsgewichte ist allerdings erschwert, wenn Suppressoreffekte vorliegen. Beispielsweise kann ein Merkmal durchaus einen bedeutsamen Zusammenhang zum Kriterium aufweisen, obwohl es keinen wesentlichen Beitrag zur Regressionsfunktion leistet.

Für die vorliegende Untersuchung war jedoch von Interesse welche Merkmale unmittelbar mit dem objektiven Wahrheitsstatus der Aussagen bzw. dem subjektiven Glaubhaftigkeitsurteil korrelierten. Dadurch sollten potentielle Diskrepanzen zwischen den ökologischen Validitäten und ihrer Nutzung für die Glaubhaftigkeitsattribution aufgedeckt werden. Zudem wird keineswegs davon auszugegangen, dass die 17 ARJS-STV-G-Merkmale trennscharf sind. Vielmehr werden beispielsweise hohe Interkorrelationen zwischen dem Merkmal Details

einerseits sowie räumlichen und zeitlichen Details erwartet. Werden solche Redundanzen herauspartialisiert gehen möglicherweise relevante Informationen verloren. Daher wurden für die nachfolgende Ergebnisdarstellung die punktbiserialen Korrelationen einander vergleichend gegenübergestellt. Die Befunde sind in Tabelle 5.6 zusammengefasst. Die anhand der multiplen Regressionsanalysen ermittelten standardisierten Beta-Koeffizienten sind zusätzlich getrennt für die freien Berichte und die Interviews im Anhang 5.a bis 5.d dokumentiert.

Freie Berichte

Für die freien Berichte ergab sich ein Zusammenhang von $r(174) = .15$, $p = .050$, zwischen dem objektiven Wahrheitsstatus und dem subjektiven Glaubhaftigkeitsurteil. Die multiple Korrelation der 17 ARJS-STV-G-Merkmale mit dem objektiven Wahrheitsstatus erklärte 17% der Varianz, adjustiertes $R^2 = .08$, $F(17,158) = 1.89$, $p = .023$. Für die Glaubhaftigkeitsattribution, ergab sich eine Varianzaufklärung von 47%, adjustiertes $R^2 = .41$, $F(17,158) = 8.11$, $p < .001$.

Tabelle 5.6 ist zu entnehmen, dass neun ARJS-STV-G-Merkmale bedeutsame Zusammenhänge zum objektiven Wahrheitsstatus aufwiesen. Dabei ergaben sich Effektstärken geringer bis mittlerer Größenordnung. Das subjektive Glaubhaftigkeitsurteil korrelierte signifikant mit der Einschätzung von 16 ARJS-STV-G-Merkmalen. Es zeigten sich geringe bis starke Effekte.

Die drei Merkmalskategorien klärten 7% der Varianz hinsichtlich des objektiven Wahrheitsstatus auf, adjustiertes $R^2 = .05$, $F(3,172) = 4.29$, $p = .006$. Für das subjektive Glaubhaftigkeitsurteil ergab sich hingegen eine Varianzaufklärung von 48%, adjustiertes $R^2 = .47$, $F(3,172) = 52.60$, $p < .001$. Die Korrelationen zwischen den drei Merkmalskategorien einerseits und dem objektiven Wahrheitsstatus sowie dem subjektiven Glaubhaftigkeitsurteil andererseits sind ebenfalls in Tabelle 5.6 aufgeführt. Die Effektstärken der ökologischen Validitäten lagen im geringen bis mittleren Bereich. Hingegen zeigten sich starke Zusammenhänge zur Glaubhaftigkeitsattribution.

Tabelle 5.6

Korrelationen (r) zwischen den 17 ARJS-STV-G-Merkmalen mit dem objektiven Wahrheitsstatus und der Glaubhaftigkeitsattribution

	Freie Berichte		Interviews	
Merkmale	r_{obj}	r_{subj}	r_{obj}	r_{subj}
Negative Äußerungen über sich selbst	**.19**	.14	.12	**.24**
Spontane Korrekturen	.01	.10	**.15**	.09
Erinnerungslücken zugeben	.03	**.15**	.04	**.24**
Komplikationen	.09	**.28**	**.17**	**.33**
Ungewöhnliche Details	**.20**	**.26**	**.26**	**.31**
Überflüssige Details	**.19**	**.24**	**.31**	**.38**
Nonverbale und verbale Interaktionen	.14	**.46**	**.28**	**.34**
Emotionen und Gefühle	**.19**	**.36**	**.18**	**.36**
Sensorische Eindrücke	.06	**.17**	-.01	**.23**
Gedächtnisprozesse und Memorieren	.11	**.25**	.15	**.40**
Gedanken	**.18**	**.34**	**.19**	**.47**
Räumliche Details	.01	**.27**	.11	**.29**
Zeitliche Details	**.15**	**.34**	.10	**.18**
Details	**.26**	**.53**	**.29**	**.52**
Persönliche Bedeutsamkeit	.03	**.29**	.06	**.36**
Klarheit und Lebendigkeit	**.24**	**.55**	**.27**	**.58**
Realismus und logische Struktur	**.20**	**.44**	**.33**	**.56**
Kategorie 1	**.18**	**.43**	**.27**	**.50**
Kategorie 2	**.17**	**.62**	**.27**	**.63**
Kategorie 3	**.23**	**.48**	**.34**	**.63**

Anm. N = 176 freie Berichte; N = 176 Interviews. r_{obj} = Zusammenhänge zwischen den ARJS-STV-G-Merkmalen und dem objektiven Wahrheitsstatus; r_{subj} = Zusammenhänge zwischen den ARJS-STV-G-Merkmalen und dem subjektiven Glaubhaftigkeitsurteil. Positive Werte signieren eine höhere Ausprägung der Merkmale in wahren bzw. als glaubhaft bewerteten Berichten.

Interviews

Für die Interviews war mit $r = .31$, $p < .001$, ein deutlicher Zusammenhang zwischen dem objektiven Wahrheitsstatus und dem subjektiven Glaubhaftigkeitsurteil nachweisbar. Die multiple Regression der 17 ARJS-STV-G-Merkmale auf den objektiven Wahrheitsstatus erzielte eine Varianzaufklärung von 26%, adjustiertes $R^2 = .18$, $F(17,158) = 3.21$, $p < .001$. Für die dichotome Glaubhaftigkeitsbeurteilung ergab sich eine Varianzaufklärung von 47%, adjustiertes $R^2 = .42$, $F(17,158) = 8.31$, $p < .001$.

Tabelle 5.6 zeigt, dass zehn Merkmale signifikant mit dem Wahrheitsstatus korrelierten. Die Effektstärken verwiesen auf geringe bis starke Zusammenhänge. Hingegen korrelierten alle Merkmale bis auf die spontanen Korrekturen signifikant mit dem dichotomen Glaubhaftigkeitsurteil. Dabei waren mittlere bis starke Effekte nachweisbar.

Die drei Merkmalskategorien erklärten nur 14% der Varianz hinsichtlich des objektiven Wahrheitsstatus, adjustiertes $R^2 = .12$, $F(3,172) = 9.15$, $p < .001$. Für das subjektive Glaubhaftigkeitsurteil wurde eine Varianzaufklärung von 51% erzielt, adjustiertes $R^2 = .50$, $F(3,172) = 59.69$, $p < .001$. Alle drei Merkmalskategorien wiesen sowohl zum objektiven Wahrheitsstatus als auch zur Glaubhaftigkeitsattribution bedeutsame Zusammenhänge auf (vgl. Tabelle 5.6). Hinsichtlich des objektiven Wahrheitsstatus waren Effekte mittlerer Größenordnung festzustellen. Hinsichtlich des subjektiven Glaubhaftigkeitsurteils ergaben sich starke Effekte.

Unterschiede zwischen den Raterinnen

Die vorliegende Untersuchung zielte darauf ab, die Effekte der ARJS-STV-G-Anleitung auf die Urteilsrichtigkeit und -neigung zu überprüfen. Daher waren die Beurteilungen mehrerer Raterinnen zusammengefasst worden. Unterschiede zwischen den Urteilen verschiedener Raterinnen waren nicht untersuchungsrelevant. Für den interessierten Leser wurden jedoch im Anhang

5.e Kennwerte zur Urteilsgüte und -neigung der einzelnen Raterinnen dokumentiert. So sind die prozentualen Urteilsrichtigkeiten und der Prozentsatz der als glaubhaft bewerteten Aussagen getrennt für jede Raterin aufgeführt. Als weitere Indikatoren der Urteilsgüte bzw. –neigung wurden A' bzw. B'' abgeleitet.

Allerdings ist darauf hinzuweisen, dass jede Raterin unterschiedliche Aussagen vor und nach der ARJS-STV-G-Anleitung bearbeitete. Obwohl die Aussagen zufällig den Sets zugeordnet und vorher nach Aussagelänge gematcht worden waren, lassen sich Veränderungen in der Urteilsgüte und –neigung einzelner Raterinnen nicht eindeutig auf die Anleitung zu den Glaubhaftigkeitsmerkmalen zurückführen. Zudem beurteilte ein Teil der Raterinnen unterschiedliche Aussagen zum selben Zeitpunkt. Demnach sind auch Unterschiede zwischen den Raterinnen nicht zwangsläufig personenbedingt. Vielmehr könnten auch Unterschiede in den beurteilten Aussagen sowohl zur intra- als auch zur interindividuellen Variabilität beigetragen haben. Auf inferenzstatistische Vergleiche wurde daher verzichtet.

Diskussion

Die vorliegende Untersuchung zielte darauf ab, Effekte einer ARJS-STV-G-Anleitung auf die Urteilsgüte, -neigung und –sicherheit zu überprüfen. Als Stimulusmaterial wurden 176 freie Berichte und 176 Interviews verwendet, die entweder unvorbereitet oder vorbereitet formuliert worden waren. Insgesamt acht Beurteilerinnen bearbeiteten diese transkribierten Aussagen. Im Rahmen eines Within-Subjects-Designs erfolgte die Beurteilung zunächst naiv und anschließend ARJS-STV-G-angeleitet.

Die Befunde zu den naiven Beurteilungen werden zunächst dem bisherigen Forschungsstand zur Entdeckung von Täuschung durch Laien gegenübergestellt. Danach werden die Effekte der ARJS-STV-G-Anleitung auf die Urteilsgüte, -neigung und -sicherheit diskutiert. Dabei wird auch auf die Validität der 17 Merkmale und ihre Bedeutung für die Glaubhaftigkeitsbeurteilung eingegangen.

Des Weiteren wird herausgearbeitet, wie sich die Vorbereitung von Aussagen auf die Entscheidungsrichtigkeit auswirkt. Schließlich werden praktische Implikationen der vorliegenden Untersuchungsbefunde aufgezeigt.

Naive Beurteilungen

Zunächst wurde angenommen, dass Laien über eine geringfügig überzufällige Fähigkeit verfügen, wahre und erfundene Aussagen zu unterscheiden (Hypothese 1a). Für das verwendete Stimulusmaterial ließ sich allerdings keine überzufällige Urteilsrichtigkeit der naiven Beurteilerinnen nachweisen. Dies galt sowohl für die freien Berichte (50.6%) als auch für die Interviews (51.1%). Die Abweichungen zu der von C. F. Bond und DePaulo (2006) berichteten gewichteten mittleren Urteilsrichtigkeit von 53.4% erscheinen jedoch vernachlässigbar. Zudem argumentierten C. F. Bond und DePaulo, dass die Urteilsgüte von Laien vielversprechender erscheint, wenn sie über die standardisierte Mittelwertsdifferenz quantifiziert wird. So ermittelten die Autoren für Studien mit mehrstufiger Beurteilungsgrundlage eine gewichtete standardisierte Mittelwertsdifferenz von $\underline{d} = 0.35$. Für die vorliegende Untersuchung fielen die standardisierten Mittelwertsdifferenzen jedoch ebenfalls gering aus ($\underline{d} = 0.09$ für die freien Berichte; $\underline{d} = 0.12$ für die Interviews). Sowohl evolutionstheoretische Überlegungen (vgl. Kraut, 1980) als auch ein doppelter Standard bei der Glaubhaftigkeitsbeurteilung (vgl. C. F. Bond & DePaulo, 2006) können diese Befunde erklären.

Bei wahren Aussagen wurde eine höhere prozentuale Urteilsrichtigkeit erzielt als bei erfundenen. Dies war auf die erwartungsgemäße Tendenz der Beurteilerinnen zurückzuführen, Aussagen als glaubhaft einzuschätzen (vgl. Hypothese 1b). Der Wahrheitsbias war sehr stark ausgeprägt. So berichteten C. F. Bond und DePaulo (2006), dass sich die Beurteiler in durchschnittlich 55.0% der Fälle für die Glaubhaftigkeit der Aussagen entschieden. Hingegen wurden in der vorliegenden Untersuchung 63.1% der freien Berichte und 70.5% der Interviews

als glaubhaft beurteilt. Dabei ist zu beachten, dass für die nachträgliche Dichotomisierung unentschiedene Urteile den als glaubhaft bewerteten Aussagen zugeordnet wurden. Dies könnte zu einer Überschätzung der liberalen Urteilsneigung geführt haben. Zusatzanalysen zeigten jedoch, dass auch unter Ausschluss der unentschiedenen Urteile noch 59.4% der freien Berichte und 69.1% der Interviews als glaubhaft eingeschätzt wurden (vgl. Tabelle 5.5). Der Wahrheitsbias war damit immer noch deutlich zu erkennen. Die vorliegenden Befunde stehen damit im Einklang mit der Vermutung, dass kognitive Heuristiken und soziale Konversationsregeln die Glaubhaftigkeitsurteile von Laien beeinflussen (z.B. O'Sullivan, 2003).

Schließlich waren sich die Beurteilerinnen trotz der zufälligen Urteilsrichtigkeit überdurchschnittlich sicher in ihren Entscheidungen. Dies unterstützt die Hypothese, dass die naiven Beurteilerinnen ihre Urteilsrichtigkeit überschätzten (Hypothese 1c). Analog zu den Befunden von DePaulo et al. (1997) sowie Aamodt und Custer (2006) war kein bedeutsamer Zusammenhang zwischen der Urteilsrichtigkeit und der subjektiven Sicherheit festzustellen. Dies lässt sich nach DePaulo et al. (1997) dadurch zu erklären, dass bei der naiven Glaubhaftigkeitsbeurteilung nicht nur richtige, sondern auch falsche und irrelevante Hinweise beachtet wurden.

Zusammenfassend waren die naiven Beurteilerinnen in der vorliegenden Untersuchung nicht dazu in der Lage wahre und erfundene Aussagen zu unterscheiden, unterlagen einem Wahrheitsbias und überschätzten ihre Urteilsgüte. Bisherige metaanalytische Befunde zur Glaubhaftigkeitsbeurteilung durch Laien wurden damit weitestgehend repliziert.

ARJS-STV-G-geleitete Beurteilungen

Urteilsgüte

Nachdem die Beurteilerinnen einen Teil der Aussagen naiv beurteilt hatten, erhielten sie eine Anleitung zu den ARJS-STV-G-Merkmalen. Dabei handelt es sich

um eine kurze Trainingsversion theoretisch fundierter und empirisch validierter inhaltlicher Aussagemerkmale (z.B. Barnier et al., 2005; Sporer, 1998; Sporer & Burghardt, 2004; Sporer & Walther, 2006). Daher wurde für die ARJS-STV-G-angeleiteten Beurteilungen eine bessere Urteilsgüte postuliert als für die naiven Beurteilungen (Hypothese 2a).

Bei den Interviews wurden sowohl für die mehrstufigen Glaubhaftigkeitsurteile als auch für die dichotome Entscheidungsrichtigkeit erwartungsgemäße Befunde nachgewiesen. So zeigte sich die postulierte Wechselwirkung zwischen der Anleitung und dem Wahrheitsstatus auf die subjektiven Glaubhaftigkeitsurteile. Dabei ergab sich ein Effekt mittlerer Größenordnung. Die naiven Beurteilerinnen hingegen waren nicht dazu in der Lage, wahre und erfundene Aussagen zu unterscheiden. Erst nachdem die Beurteilerinnen über die ARJS-STV-G informiert worden waren, erzielten wahre Aussagen höhere Bewertungen als erfundene. Dies war auf eine veränderte Einschätzung der wahren Aussagen zurückzuführen. Auch die Analysen zur prozentualen Urteilsrichtigkeit demonstrierten, dass die wahren Aussagen mit der ARJS-STV-G-Anleitung häufiger richtig eingeschätzt wurden als ohne die Anleitung. Für die ARJS-STV-G-angeleiteten Beurteilungen ergab sich eine äußerst hohe Trefferrate von 94.3%, während sie für die naiven Beurteilungen bei 71.6% lag. Aufgrund der liberalen Urteilsneigung war für die ARJS-STV-G-angeleiteten Beurteilungen eine zufällige Trefferrate von 82.4% zu erwarten, für die naiven von 70.5%. Daher ist nur die Trefferrate der angeleiteten Beurteilerinnen als überzufällig zu werten. Die ARJS-STV-G schien demnach bei der Einschätzung wahrer Aussagen hilfreich zu sein. Allerdings blieb die Urteilsgüte bei erfundenen Aussagen gering.

Für die freien Berichte verfehlte die Wechselwirkung zwischen der Anleitung und dem Wahrheitsstatus auf die subjektiven Glaubhaftigkeitsurteile hingegen knapp statistische Signifikanz. Auch hinsichtlich der prozentualen Urteilsrichtigkeit ließ sich keine statistisch bedeutsame Interaktion zwischen dem Wahrheitsstatus

und der Anleitung nachweisen. Dennoch waren für beide Analysen Effektstärken geringer Größenordnung in erwarteter Richtung festzustellen. Zudem fielen für die subjektiven Glaubhaftigkeitsurteile die einfachen Haupteffekte erwartungsgemäß aus. Analog zu den Beurteilungen der Interviews zeigte sich, dass wahre Aussagen mit der ARJS-STV-G-Anleitung signifikant höhere Beurteilungen erzielten als ohne die Anleitung.

Bei vorangegangenen Analysen (vgl. Studie 4) wurden keine signifikanten Unterschiede in der Validität der ARJS-STV-G-Merkmale für die freien Berichte und die Interviews festgestellt. Es zeigte sich jedoch ein Haupteffekt der Aussageform, der auf eine höhere Qualität der Interviews im Vergleich zu den freien Berichten verwies. Dies könnte erklären, warum für die Interviews im Gegensatz zu den freien Berichten ein positiver Effekt der Anleitung nachweisbar war. Möglicherweise profitierten die Beurteilerinnen erst dann von der Anleitung, wenn die Glaubhaftigkeitsmerkmale sehr häufig im Stimulusmaterial vorzufinden waren. Zudem ist zu beachten, dass für die Entscheidungsrichtigkeit bei den freien Berichten differentielle Effekte der Vorbereitung auf die ARJS-STV-G-angeleitete Einschätzung wahrer und erfundener Aussagen wirksam waren.

Die vorliegenden Befunde verweisen darauf, dass die ARJS-STV-G die korrekte Einschätzung wahrer Aussagen unterstützen können. Die ARJS-STV-G-Merkmale sind als Glaubhaftigkeitsmerkmale aufzufassen, d.h. ihr Vorhandensein stärkt die Hypothese eines persönlichen Erlebnisbezugs. Wahre Aussagen, bei denen die Merkmale in hoher Quantität und/oder Qualität vorliegen, sollten mit der ARJS-STV-G-Anleitung recht eindeutig zu beurteilen sein. Für die Einschätzung erfundener Aussagen liefert die ARJS-STV-G hingegen nur indirekt Hilfestellung. So kann das Fehlen der Merkmale zwar nicht die Glaubhaftigkeitshypothese stärken, doch bedeutet das nicht zwangsläufig, dass eine Aussage erfunden ist. Es wird durchaus eingeräumt, dass die ARJS-STV-G-Merkmale in einer geringeren Quantität und/oder Qualität auch bei erfundenen Aussagen vorzufinden sind. In solchen Fällen folgen die Beurteilerinnen vermutlich doch wieder ihrer

Intuition, um eine Entscheidung zu treffen. Daher erscheint es nachvollziehbar, dass die Beurteilerinnen nur bei der Beurteilung wahrer Interviews von der ARJS-STV-G-Anleitung profitierten.

Bisherige Studien zur Trainierbarkeit der Urteilsgüte anhand von CBCA-Merkmalen berichteten widersprüchliche Befunde. Wurde die Validität der Glaubhaftigkeitsmerkmale für das verwendete Stimulusmaterial nicht überprüft (Köhnken, 1987a) oder nicht belegt (Akehurst et al., 2004; Ruby & Brigham, 1998) zeigten sich keine positiven Effekte auf die Urteilsgüte. Hingegen berichteten Landry und Brigham (1992) eine höhere prozentuale Urteilsrichtigkeit für CBCA-trainierte im Vergleich zu naiven Beurteilern (allerdings nur bei Videoaufnahmen, nicht bei Transkripten). Dies war im Einklang mit den Befunden der vorliegenden Untersuchung auf eine verbesserte Einschätzung wahrer Aussagen zurückzuführen. Die von Steller (1989) berichteten positiven Trainingseffekte ergaben sich hingegen aus einer verbesserten Einschätzung wahrer und erfundener Aussagen.

Sporer (1998) wiederum konnte zeigen, dass zwei ARJS-Beurteiler eine höhere prozentuale Urteilsrichtigkeit aufwiesen als zwei naive Beurteiler. Im Gegensatz dazu wiesen Studien, in denen die Beurteiler nur über einen Teil der ARJS-Merkmale informiert wurden, bislang keine positiven Trainingseffekte nach. Sporer et al. (2002) untersuchten die Effekte einer kurzen Anleitung zu neun ARJS-Merkmalen auf die Urteilsgüte. Im Vergleich zu einer naiven Kontrollgruppe wiesen die angeleiteten Beurteiler zwar augenscheinlich eine höhere Entscheidungsrichtigkeit auf, die Unterschiede waren jedoch nicht signifikant. Sporer und Masip (2007) wiederum fanden keine Unterschiede in der Urteilsgüte naiver und anhand der 17 Merkmale der ARJS-Kurzform angeleiteter Beurteiler. Dies führten die Autoren auf die Konstruktion des Stimulusmaterials zurück. Eine weitere Beurteilergruppe wurde über vier Merkmale der ARJS-Kurzform informiert und erzielte eine höhere prozentuale Urteilsrichtigkeit als die Kontrollgruppe bei erfundenen Aussagen. Dies war jedoch auf eine veränderte Urteilsneigung

zurückzuführen. Insgesamt fiel die Urteilsgüte nicht besser aus als bei der naiven Kontrollgruppe. Sporer und Masip (2007) beschränkten die Anleitung auf schriftliche Informationen zu den Glaubhaftigkeitsmerkmalen. Im Gegensatz dazu erhielten die Beurteilerinnen der vorliegenden Untersuchung eine kurze Schulung zu der ARJS-STV-G. Dabei hatten sie Gelegenheit, die Anwendung anhand von Beispielsätzen einzuüben und Verständnisschwierigkeiten zu klären.

Zusammenfassend legen die Befunde von Steller (1989) und Landry und Brigham (1992) nahe, dass selbst eine kurze Einführung zu Glaubhaftigkeitsmerkmalen (90 und 45 Minuten respektive) positive Effekte auf die Urteilsgüte haben kann. Dies scheint nach den vorliegenden Befunden auch für eine ARJS-STV-G-Anleitung zu gelten. Die relevanten Informationen schriftlich zu vermitteln, reicht dabei vermutlich nicht aus (vgl. Sporer & Masip, 2007). Vielmehr sollten die Beurteilerinnen Gelegenheit haben, die Anwendung der Merkmale einzuüben und zu diskutieren. Ähnliche Anforderungen an Trainingsstudien wurden von Frank und Feeley (2003) formuliert. Sie argumentierten, dass Schulungen mindestens 50 Minuten beanspruchen und Gelegenheit zum Üben sowie Rückmeldungen hinsichtlich der Richtigkeit von Glaubhaftigkeitsurteilen gewähren sollten.

Urteilsneigung

Da die ARJS-STV-G ausschließlich Glaubhaftigkeitsmerkmale beinhaltet, wurde eine Verstärkung der liberalen Urteilsneigung infolge der ARJS-STV-G-Anleitung erwartet (Hypothese 2b). Diese Annahme ließ sich erneut für die Interviews bestätigen. So zeigte sich erwartungsgemäß, dass mit der Anleitung höhere Glaubhaftigkeitsbeurteilungen vergeben wurden als ohne die Anleitung. Für die freien Berichte war hingegen kein statistisch signifikanter Haupteffekt der Anleitung nachweisbar.

Leider wurde nur selten überprüft, ob ein Training zu inhaltlichen Glaubhaftigkeitsmerkmalen liberale Urteilstendenzen verstärkt. Die Befunde von Landry und Brigham (1992) stehen jedoch im Einklang mit den vorliegenden

Untersuchungsbefunden. Die Autoren fanden, dass sich eine CBCA-Trainingsgruppe signifikant häufiger für die Glaubhaftigkeit der Aussagen entschied als eine naive Kontrollgruppe. Köhnken (1987a) verglich die Urteilsgüte einer Kontrollgruppe mit der von drei Trainingsgruppen, von denen eine über CBCA-Merkmale informiert worden war. Für jede Untersuchungsgruppe wurden die Raten der Treffer und korrekten Zurückweisungen berichtet. Anhand dieser Angaben ließen sich signalentdeckungstheoretische Indizes zur Urteilsneigung ableiten (vgl. Tabelle 5.3). Die Werte tendierten in die erwartete Richtung. Für die Kontrollgruppe ergab sich eine Urteilsneigung von B" = -.57, für die CBCA-Trainingsgruppe von B" = -.73. Allerdings zeigten sich keine signifikanten Unterschiede zwischen allen vier Untersuchungsgruppen in der Entscheidungsrichtigkeit für wahre und erfundene Aussagen.

Sporer und Bursch (1996) wiederum ließen Beurteiler dasselbe Stimulusmaterial zunächst naiv und dann anhand von CBCA- und RM-Merkmalen beurteilen. Während die naiven Beurteilungen ausgewogen waren, war nach der Information zu den Glaubhaftigkeitsmerkmalen ein leichter Wahrheitsbias zu erkennen. Bisherige Untersuchungen zu den ARJS fanden hingegen keine trainingsbedingte Verstärkung des Wahrheitsbias (Sporer, 1998; Sporer et al., 2002). Überraschenderweise berichteten Sporer und Masip (2007) einen gegensätzlichen Effekt. Die naive Kontrollgruppe und die anhand von 17 Merkmalen der ARJS-Kurzform angeleitete Gruppe zeigten eine liberale Urteilsneigung. Hingegen wählten die Beurteiler ein ausgewogenes Entscheidungskriterium, wenn sie lediglich über vier Glaubhaftigkeitsmerkmale der Kurzform informiert worden waren. Die Autoren diskutierten, dass die Beurteiler insgesamt kaum Glaubhaftigkeitsmerkmale in den Aussagen fanden, weil sie nur auf vier Merkmale achteten. Infolgedessen erscheint es nachvollziehbar, dass keine Tendenz vorlag Aussagen als glaubhaft zu beurteilen. Umgekehrt fanden Beurteiler, die auf 17 Aussagemerkmale achten sollten, vergleichsweise häufig Glaubhaftigkeitsmerkmale. Allerdings erlaubten die

meisten Merkmale für das verwendete Stimulusmaterial keine valide Unterscheidung wahrer und erfundener Aussagen. Die anhand der 17 Merkmale angeleitete Gruppe wurde demnach auf überwiegend für die vorliegenden Aussagen nicht valide Glaubhaftigkeitsmerkmale aufmerksam gemacht. Dies resultierte in einer liberalen Urteilsneigung, ohne die Urteilsgüte zu verbessern.

Die Argumentation von Sporer und Masip (2007) lässt sich auch auf die vorliegenden Untersuchungsbefunde anwenden. So zeigte sich eine höhere Qualität der Interviews im Vergleich zu den freien Berichten (vgl. Studie 4). Dies könnte erklären, warum für die Interviews im Gegensatz zu den freien Berichten eine Verstärkung der Urteilsneigung festgestellt wurde. Möglicherweise verändert sich die Urteilsneigung erst, wenn die Glaubhaftigkeitsmerkmale sehr häufig im Stimulusmaterial vorzufinden sind.

Zusammenfassend zeigte sich erwartungsgemäß, dass eine ARJS-STV-G-Anleitung liberale Urteilstendenzen verstärken kann. Daher könnte eine solche Anleitung zu Glaubhaftigkeitsmerkmalen vor allem Polizisten zu empfehlen sein (vgl. Garrido et al., 2004). So wurde für diese Berufsgruppe wiederholt ein Lügenbias festgestellt (Ekman et al., 1999; Garrido et al., 2004; Meissner & Kassin, 2002; Porter et al., 2000; Vrij et al., 2007), der dadurch kompensiert werden könnte. Allerdings bleibt ungeklärt, unter welchen Bedingungen das Wissen um Glaubhaftigkeitsmerkmale die Urteilsneigung beeinflusst bzw. wann sie stabil bleibt. Dies erfordert weitere Forschungsbemühungen. Dabei wäre es möglicherweise sinnvoll, die inhaltliche Qualität von Aussagen gezielt zu manipulieren und Effekte auf die Urteilsneigung zu untersuchen. Zudem wäre es interessant zu überprüfen, ob eine konservative Urteilsneigung resultiert, wenn Lügenmerkmale trainiert werden (vgl. Kassin & Fong, 1999).

Urteilssicherheit

Aussagen mit einer hohen inhaltlichen Qualität sollten einen persönlichen Erlebnisbezug aufweisen (vgl. Sporer, 1996/1998/2004). Durch die ARJS-STV-G-Anleitung wurden die Beurteilerinnen angewiesen, solche Aussagen als glaubhaft

zu beurteilen. Umgekehrt konnten sie die ARJS-STV-G nutzen, um ihre Entscheidungen wissenschaftlich zu rechtfertigen. Dies sollte sich wiederum positiv auf ihre subjektive Sicherheit auswirken. Daher wurde für die ARJS-STV-G-geleiteten Beurteilungen ein bedeutsamer Zusammenhang zwischen der subjektiven Sicherheit und der Entscheidungsrichtigkeit erwartet (Hypothese 2c). Für die Beurteilungen der freien Berichte war ein hypothesenkonform positiver Zusammenhang nachweisbar. Bei den Interviews gingen richtige Entscheidungen hingegen nur tendenziell mit einer höheren subjektiven Sicherheit einher, wobei die subjektive Sicherheit sowohl bei richtigen als auch bei falschen Entscheidungen eher hoch war.

In Trainingsstudien finden sich nur selten Angaben zu der Korrelation zwischen der Urteilsrichtigkeit und –sicherheit für die einzelnen Untersuchungsgruppen. Diese Korrelation birgt in der Tat eine gewisse Redundanz mit der Korrelation zwischen dem Glaubhaftigkeitsurteil und dem Wahrheitsstatus. So werden Glaubhaftigkeitsurteile oft anhand mehrstufiger Skalen erhoben, obwohl der objektive Wahrheitsstatus der Aussagen dichotom ausgeprägt ist. Mehrstufige Urteile spiegeln demnach nicht nur die Glaubhaftigkeitsattribution wider, sondern enthalten auch Informationen zur subjektiven Sicherheit der Beurteiler (Köhnken, 1987a). Akehurst et al. (2004) sowie Köhnken (1987a) berichteten jedoch entsprechende Zusammenhänge zwischen der Urteilsgüte und der getrennt erfassten subjektiven Sicherheit. So fand Köhnken (1987a) für die Gruppe der CBCA-trainierten Beurteiler einen bedeutsamen Zusammenhang von $\underline{r}$ = .26, der für die naive Kontrollgruppe ($\underline{r}$ = .04) nicht festzustellen war. Akehurst et al. (2004) wiesen hingegen keine Wechselwirkung zwischen dem Erhebungszeitpunkt (vor und nach der CBCA-Information) und der Entscheidungsrichtigkeit auf die subjektive Sicherheit nach. Zudem zeigte sich kein Haupteffekt des Erhebungszeitpunkts auf die subjektive Sicherheit. Die Autoren konnten allerdings auch keine Trainingseffekte hinsichtlich der Urteilsgüte nachweisen.

Zusammenfassend verbesserten die ARJS-STV-G-angeleiteten Beurteilerinnen die subjektive Einschätzung ihrer Entscheidungsrichtigkeit bei den freien Berichten. Für das verwendete Stimulusmaterial waren die meisten ARJS-STV-G-Merkmale valide Indikatoren des objektiven Wahrheitsstatus und nur wenige Merkmale erwiesen sich als irrelevant (vgl. Studie 4). Infolgedessen kam es möglicherweise zu einer besseren Kallibrierung der subjektiven Sicherheit (vgl. DePaulo et al., 1997).

Brunswiksche Linsenmodellanalyse

Um die Validität der ARJS-STV-G-Merkmale und ihre Zusammenhänge zur Glaubhaftigkeitsattribution vergleichend zu betrachten, wurde eine Brunswiksche Linsenmodellanalyse durchgeführt. Es zeigte sich erwartungsgemäß eine starke Korrespondenz zwischen den Merkmalsbewertungen und dem subjektiven Glaubhaftigkeitsurteil (Hypothese 5). Die Beurteiler beachteten fast alle 17 Merkmale bei der Glaubhaftigkeitsattribution. So zeigten sich bei den freien Berichten für 15, bei den Interviews für 16 Merkmale signifikante Korrelationen zum Glaubhaftigkeitsurteil. Dabei gingen hohe Merkmalsausprägungen instruktionsgemäß mit hohen Glaubhaftigkeitsbeurteilungen einher. Zudem wiesen alle drei Merkmalskategorien starke Zusammenhänge zum subjektiven Glaubhaftigkeitsurteil auf. Die Beurteilerinnen verwendeten demnach nicht die vorgegebenen Gewichtungsregeln. Allerdings zeigten sich auch kaum Unterschiede in den ökologischen Validitäten der drei Merkmalskategorien. Die Gewichtungsregeln wurden von Sporer (1996/1998/2004) aufgrund theoretischer Überlegungen sowie erster Ergebnisse zu den ARJS-Merkmalen formuliert. Sämtliche ARJS-Merkmale zu vermitteln erfordert eine intensive und zeitlich aufwändige Beurteilerschulung (vgl. Studie 3). Im Gegensatz dazu kann eine kurze Anleitung zu den 17 ARJS-STV-G-Merkmalen lediglich ein Grundverständnis der Glaubhaftigkeitsmerkmale leisten. So bleibt zu spekulieren, dass ARJS-STV-G-angeleitete Beurteilerinnen die Merkmale im Einzelfall anders anwenden als ARJS-Experten.

Der Befund eines stärkeren Zusammenhangs der Glaubhaftigkeitsmerkmale zu dem subjektiven Urteil im Vergleich zum tatsächlichen Wahrheitsstatus steht im Einklang mit anderen Brunswikschen Linsenmodellanalysen (vgl. Fiedler, 1989a, 1989b; Sporer & Küpper, 1995; Reinhard et al., 2002; Cramer, 2005). Insgesamt erzielten die Glaubhaftigkeitsmerkmale eine Varianzaufklärung von 47% hinsichtlich der Glaubhaftigkeitsurteile für die freien Berichte und die Interviews. Für die Glaubhaftigkeitsbeurteilungen anhand der Langform der ARJS wurde eine höhere Varianzaufklärung von 79% berichtet (Sporer, 1998).

Es bedarf weiterer Forschung, um mögliche Unterschiede im Verständnis der ARJS und der ARJS-STV-G-Merkmale aufzudecken. Solche Unterschiede könnten gegebenenfalls auch eine Revision der ARJS-STV-G-Gewichtungsregeln sinnvoll erscheinen lassen.

Moderierende Effekte der Vorbereitung

Gemäß der Befunde von C. F. Bond und DePaulo (2006) wurde angenommen, dass vorbereitete Aussagen schwieriger einzuschätzen sind als unvorbereitete (Hypothese 3). Allerdings wurden weder für die freien Berichte noch für die Interviews Haupteffekte der Vorbereitung auf die Urteilsrichtigkeit festgestellt. Bei den freien Berichten zeigte sich jedoch eine Wechselwirkung zwischen dem Wahrheitsstatus, der Vorbereitungszeit und der Anleitung auf die Entscheidungsrichtigkeit. Für die naiven Beurteilungen ergaben sich weder bei wahren noch bei erfundenen Aussagen Unterschiede in Abhängigkeit von der Vorbereitung. Im Gegensatz dazu verschlechterten sich durch die Vorbereitung die ARJS-STV-G-angeleiteten Einschätzungen der erfundenen Aussagen, während sich die der wahren verbesserten.

Dies ist vermutlich darauf zurückzuführen, dass es den Stimuluspersonen durch die Vorbereitung gelang, ihre freien Berichte unabhängig vom objektiven Wahrheitsgehalt inhaltlich anzureichern. Diese Argumentation wird durch die

Befunde zur Validität der ARJS-STV-G unterstützt (vgl. Studie 4). So zeigte sich eine höhere Qualität vorbereiteter im Vergleich zu unvorbereiteten freien Berichten. Für die Interviews wurde hingegen kein Haupteffekt der Vorbereitung nachgewiesen. Dies ist vermutlich dadurch bedingt, dass die experimentelle Manipulation der Vorbereitung eine Woche vor der Erhebung der Interviews erfolgte.

Die naiven Beurteilerinnen bemerkten oder beachteten diese Unterschiede in der Qualität vorbereiteter und unvorbereiteter freier Berichte anscheinend nicht. Hingegen wurden sie durch die ARJS-STV-G-Anleitung für die inhaltlichen Glaubhaftigkeitsmerkmale und deren Bedeutsamkeit sensibilisiert. Die richtige Einschätzung erfundener Aussagen wurde durch Vorbereitung erschwert, die wahrer Aussagen erleichtert.

Es bleibt festzuhalten, dass vorbereitungsbedingte Unterschiede in der inhaltlichen Aussagequalität die Urteilsgüte beeinflussen können. Dies ist jedoch nicht zwangsläufig negativ zu werten. Vielmehr hängt die Wertung von der relativen Wichtigkeit der beiden Arten richtiger Entscheidungen ab. Schließlich scheint sich die richtige Bewertung wahrer Aussagen durch Vorbereitung zu verbessern.

Naive, ARJS-STV-G-geleitete und Expertenurteile

Zudem sollte die Urteilsgüte der naiven und ARJS-STV-G-angeleiteten Beurteilerinnen mit der von Experten verglichen werden. Dazu wurden die Daten von Studie 3 herangezogen, in der vier Experten dasselbe Stimulusmaterial anhand der ARJS-Langform bewerteten. Es wurde angenommen, dass diese Experten eine höhere Urteilsgüte aufweisen würden als die naiven und ARJS-STV-G-angeleiteten Beurteilerinnen der vorliegenden Untersuchung (Hypothese 4). Die prozentualen Urteilsrichtigkeiten verweisen auf hypothesenkonforme Befunde. Die naiven Beurteilerinnen bewerteten sowohl die freien Berichte als auch die Interviews auf Zufallsniveau. Hingegen wiesen die ARJS-STV-G-angeleiteten Beurteilerinnen eine überzufällige prozentuale Urteilsrichtigkeit bei der Einschätzung der Interviews auf. Allerdings waren sie nicht dazu in der Lage die

freien Berichte überzufällig richtig zu beurteilen. Nur die Expertengruppe erzielte sowohl bei den Interviews als auch bei den freien Berichten eine überzufällige Entscheidungsrichtigkeit.

Für die Glaubhaftigkeitsbeurteilung wurden 10-stufige Skalen verwendet. Dadurch war es möglich eindeutig zwischen einer Beurteilung als glaubhaft und nicht-glaubhaft zu unterscheiden. In den beiden Studien 3 und 4 waren jedoch die Einschätzungen verschiedener Beurteilergruppen von zentralem Interesse. Daher wurden die Urteile mehrerer Beurteilerinnen zusammengefasst, so dass unentscheidbare Fälle resultierten. Für die statistischen Analysen wurden Aussagen mit einem mittleren Glaubhaftigkeitsurteil von 5.5 den als glaubhaft beurteilten Aussagen zugeordnet. Studien, die den Beurteilern beispielsweise über 5- oder 9-stufige Beurteilungsskalen die Möglichkeit einräumten keine eindeutigen Entscheidung zu treffen, schlossen diese Fälle meist nachträglich von den Analysen aus (z.B. Landry & Brigham, 1992; Steller, 1989). Daher wurden auch für die vorliegende Untersuchung die prozentualen Urteilsrichtigkeiten zusätzlich unter Ausschluss der wenigen unentscheidbaren Fälle ermittelt. Die Befunde sind in Tabelle 5.5 aufgeführt. Für alle Beurteilergruppen wurden die prozentualen Urteilsrichtigkeiten dadurch augenscheinlich leicht erhöht. Zudem ließ sich nun auch für die ARJS-STV-G-angeleiteten Beurteilungen der freien Berichte eine überzufällige Urteilsrichtigkeit nachweisen.

Praktische Implikationen

Die ARJS beanspruchen einen breiten Anwendungsbereich. Die vorliegende Untersuchung demonstrierte, dass sich auch eine Anleitung anhand der ökonomischen Kurzform positiv auf die Urteilsgüte auswirken kann. Die Anleitung beanspruchte lediglich zweieinhalb Stunden und könnte daher für viele Berufsgruppen eine interessante Weiterbildungsmaßnahme darstellen. Schließlich zeigt die Täuschungsliteratur, dass beispielsweise Polizisten, Detektive, Richter und Psychologen wahre und erfundene Aussagen nicht besser

unterscheiden können als naive Studierende (Aamodt & Custer, 2006). Dennoch ist zu beachten, dass trotz beeindruckender Trefferraten (76.1% für die freien Berichte und 94.3% für die Interviews), die Rate der korrekten Zurückweisungen gering blieb (37.5% für die freien Berichte und 29.5% für die Interviews). Dies verweist auf liberale Urteilstendenzen, die durch eine ARJS-STV-G-Anleitung möglicherweise noch verstärkt werden können.

Umgekehrt könnte es jedoch auch sein, dass Personen mit einer konservativen Urteilsneigung nach einer ARJS-STV-G-Anleitung ausgewogener urteilen (vgl. Garrido et al., 2004). Zudem sind liberale Urteilstendenzen nicht zwangsläufig negativ zu bewerten. So kann ein Wahrheitsbias von Vorteil sein, wenn die Basisrate sehr hoch ist oder es besonders wichtig ist, wahre Aussagen richtig zu beurteilen. Die Effekte einer Basisratenverschiebung auf die Urteilsrichtigkeit lassen sich anhand der von Park und Levine (2001) vorgeschlagenen Formeln abschätzen. So lässt sich für die vorliegenden Untersuchungsbefunde demonstrieren, dass eine Basisratenerhöhung nicht nur die Trefferrate, sondern auch die Urteilsrichtigkeit insgesamt verbessern würde (vgl. Abbildung 5.8).

Allerdings ist die Basisrate bei rechtspsychologisch relevanten Fragestellungen nie bekannt. Umfrageuntersuchungen zur Basisrate für spezifische Delikte (z.B. Everson, Boat, Bourg & Robertson, 1996) spiegeln lediglich subjektive Einschätzungen wider. Ob solche Schätzungen der Realität entsprechen, bleibt jedoch absolut unklar. Selbst wenn die Basisraten bekannt wären, so ließe sich die Frage nach der relativen Wichtigkeit der Urteilsgüte für wahre und erfundene Aussagen wohl kaum zufriedenstellend klären. Schließlich können Fehldiagnosen jeglicher Art im rechtspsychologischen Bereich drastische unerwünschte Konsequenzen haben.

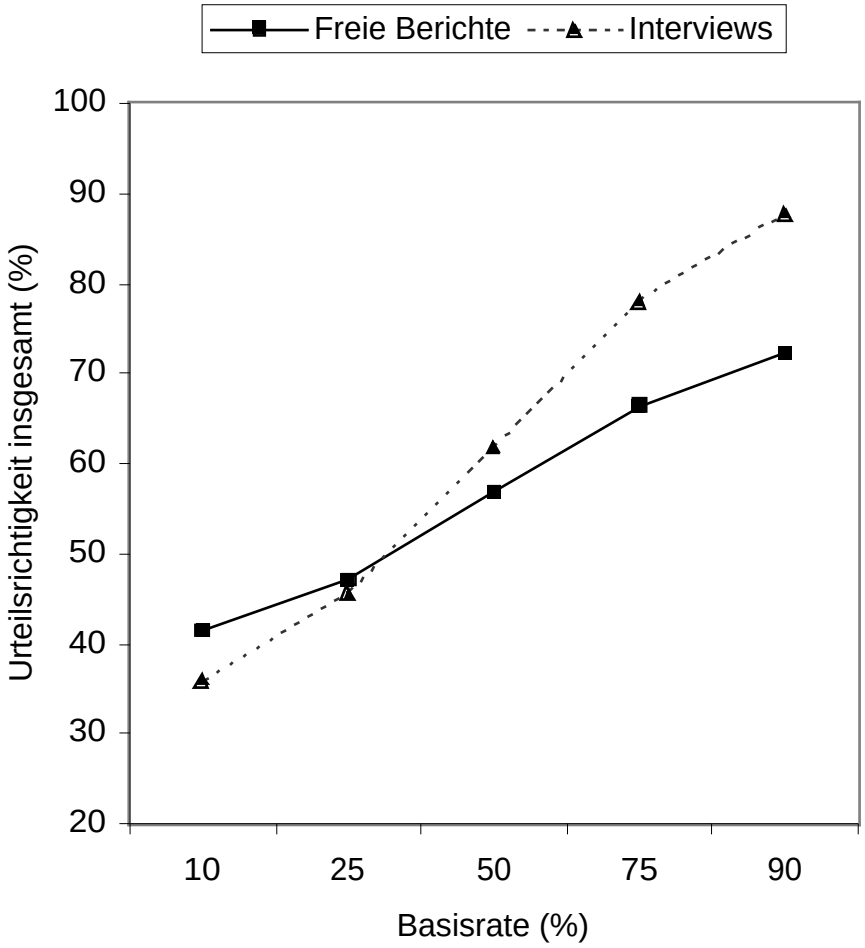


Abbildung 5.8. Abschätzung der aufgrund der vorliegenden Untersuchungsbefunde zu erwartenden prozentualen Urteilsrichtigkeit bei Basisratenverschiebungen auf 10%, 25%, 75% und 90%.

Daher bleibt das Ziel, die Beurteilungsfähigkeit insgesamt zu verbessern. Die vorliegende Untersuchung zeigte erstmals, dass ARJS-STV-G-Anleitungen dazu beitragen könnten. Weitere Untersuchungen sollten klären, ob sich die positiven Effekte auf die Urteilsgüte für anderes Stimulusmaterial und verschiedene Beurteilergruppen replizieren lassen.

Literatur

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar: A meta-analysis of individual differences in detecting deception. Forensic Examiner, 15, 6-11.

Akehurst, L., Bull, R., Vrij, A., & Köhnken, G. (2004). The effects of training professional groups and lay persons to use criteria-based content analysis to detect deception. Applied Cognitive Psychology, 18, 877-891.

Akehurst, L., Köhnken, G., Vrij, A., & Bull, R. (1996). Lay persons' and police officers' beliefs regarding deceptive behaviour. Applied Cognitive Psychology, 10, 461-471.

Barnier, A. J., Sharman, S. J., McKay, L., & Sporer, S. L. (2005). Discriminating adults' genuine, imagined, and deceptive accounts of positive and negative childhood events. Applied Cognitive Psychology, 19, 985-1001.

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. Personality and Social Psychology Review, 10, 214-234.

Bond, C. F., Jr., (in press). A few can catch a liar, sometimes. Applied Cognitive Psychology.

Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. Applied Cognitive Psychology, 19, 313-329.

Bond, G. D., Malloy, D. M., Arias, E. A., Nunn, S. N., & Thompson, L. A. (2005). Lie-biased decision making in prison. Communication Reports, 18, 9-19.

Bond, G. D., Thompson, L. A., & Malloy, D. M. (2005). Vulnerability of older adults to deception in prison and nonprison contexts. Psychology and Aging, 20, 9-19.

Bond, C. F., Jr., & Uysal, A. (2007). On lie detection “wizards”. Law and Human Behavior, 31, 109-115.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale: Lawrence Erlbaum.

Cramer, M. (2005). Can we train people to detect deception? Unpublished master’s thesis, Justus-Liebig-Universität Gießen, Germany.

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J., & Muhlenbruck, L. (1997). The accuracy-confidence correlation in the detection of deception. Personality and Social Psychology Review, 1, 346-357.

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

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. Psychological Bulletin, 129, 74-112.

DePaulo, B. M., & Pfeifer, R. L. (1986). On-the-job experience and skill at detecting deception. Journal of Applied Social Psychology, 16, 249-267.

DePaulo, B. M., Zuckerman, M., & Rosenthal, R. (1980). Humans as lie detectors. Journal of Communication, 30, 129-139.

DeTurck, M. A., & Miller, G. R. (1990). Training observers to detect deception: Effects of self-monitoring and rehearsal. Human Communication Research, 16, 603-620.

Donaldson, W. (1992). Measuring recognition memory. Journal of Experimental Psychology: General, 121, 275-277.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. Psychological Methods, 1, 170-177.

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar? American Psychologist, 46, 913-920.

Ekman, P., O'Sullivan, M., & Frank, M. G. (1999). A few can catch a liar. Psychological Science, 10, 263-266.

Everson, M. D., Boat, B. W., Bourg, S., & Robertson, K. R. (1996). Beliefs among professionals about rates of false allegations of child sexual abuse. Journal of Interpersonal Violence, 11, 541-553.

Fiedler, K. (1989a). Lügendetektion aus alltagspsychologischer Sicht. Psychologische Rundschau, 40, 127-140.

Fiedler, K. (1989b). Suggestion and credibility: Lie detection based on content-related cues. In V. A. Gheorghin, P. Netter, H. J. Eysenck, & R. Rosenthal (Eds.), Suggestion and suggestiblity: Theory and research (pp. 323-335). Berlin: Springer.

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

Frank, M. G., & Feeley, T. H. (2003). To catch a liar: Challenges for research in lie detection training. Journal of Applied Communication Research, 31, 58-75.

Garrido, E., Masip, J., & Herrero, C. (2004). Police officers' credibility judgments: Accuracy and estimated ability. International Journal of Psychology, 39, 254-275.

Granhag, P. A., Strömwall, L. A., & Landström, S. (2006). Children recalling an event repeatedly: Effects on RM and CBCA scores. Legal and Criminological Psychology, 11, 81-98.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), Syntax and semantics (pp. 41-58). New York: Academic Press.

Hammond, K. R., & Stewart, T. (Eds.).(2001). The essential Brunswik. Beginnings, explications, applications. New York: Oxford University Press.

Hartwig, M., Granhag, P. A., Strömwall, L. A., & Andersson, L. O. (2004a). Suspicious minds: Criminals' ability to detect deception. Psychology, Crime, and Law, 10, 83-95.

Hartwig, M., Granhag, P. A., Strömwall, L. A., & Kronkvist, O. (2006). Strategic use of evidence during police interviews: When training to detect deception works. Law and Human Behavior, 30, 603-619.

Hartwig, M., Granhag, P. A., Strömwall, L. A., & Vrij, A. (2004b). Police officers' lie detection accuracy: Interrogating freely versus observing video. Police Quaterly, 7, 429-456.

Johnson, M. K., Foley, M. A., Suengas, A. G., & Raye, C. L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. Journal of Experimental Psychology: General, 117, 371-376.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. Psychological Review, 88, 67-85.

Kalbfleisch, P. J. (1990). Listening for deception: The effects of medium on accuracy of deception. In R. N. Bostrom (Ed.), Listening behavior: Measurement and application (pp. 155-176). New York: Guilford Press.

Kassin, S. M., & Fong, C. T. (1999). "I'm innocent!": Effects of training on judgments of truth and deception in the interrogation room. Law and Human Behavior, 23, 499-516.

Kassin, S. M., Meissner, C. A., & Norwick, R. J. (2005). "I'd know a false confession if I saw one": A comparative study of college students and police investigators. Law and Human Behavior, 29, 211-227.

Köhnken, G. (1987a). Training police officers to detect deceptive eyewitness statements: Does it work? Social Behaviour, 2, 1-17.

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

Krahé, B., & Kundrotas, S. (1992). Glaubwürdigkeitsbeurteilung bei Vergewaltigungsanzeigen. Zeitschrift für experimentelle und angewandte Psychologie, 4, 598-620.

Kraut, R. (1980). Humans as lie detectors: Some second thoughts. Journal of Communication, 30, 209-216.

Küpper, B., & Sporer, S. L. (1995). Beurteilerübereinstimmung bei Glaubwürdigkeitsmerkmalen: Eine empirische Studie. In G. Bierbrauer (Ed.), Verfahrensgerechtigkeit: Rechtspsychologische Forschungsbeiträge für die Justizpraxis (S. 187-213). Köln: Verlag Dr. Otto Schmidt KG.

Landry, K., & Brigham, J. C. (1992). The effect of training in Criteria-Based Content Analysis on the ability to detect deception in adults. Law and Human Behavior, 16, 663-675.

Leach, A.-M., Talwar, V., Lee, K., Bala, N., & Lindsay, R. C. L. (2004). "Intuitive" lie detection of children's deception by law enforcement officials and university students. Law and Human Behavior, 28, 661-685.

Levine, T. R., & McCornack, S. A. (1991). The dark side of trust: Conceptualizing and measuring types of communicative suspicion. Communication Quaterly, 39, 325-340.

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect". Communication Monographs, 66, 125-144.

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. Thousand Oaks, CA: Sage.

Mann, S., & Vrij, A. (2006). Police officers' judgements of veracity, tenseness, cognitive load and attempted behavioural control in real-life police interviews. Psychology, Crime, and Law, 12, 307-319.

Mann, S., Vrij, A., & Bull, R. (2004). Detecting true lies: Police officers' ability to detect suspects' lies. Journal of Applied Psychology, 89, 137-149.

Mann, S., Vrij, A., & Bull, R. (2006). Looking through the eyes of an accurate lie detector. The Journal of Credibility Assessment and Witness Psychology, 7, 1-16.

Mann, S., Vrij, A., Fisher, R. P., & Robinson, M. (2007, published online). See no lies, hear no lies: Differences in discrimination accuracy and response bias when watching or listening to police suspect interviews. Applied Cognitive Psychology.

Masip, J., Alonso, H., Garrido, E., & Antón, C. (2005). Generalized communicative suspicion among police officers: Accounting for the investigator bias effect. Journal of Applied Social Psychology, 35, 1046-1066.

Masip, J., Garrido, E., & Herrero, C. (2003). When did you conclude she was lying? The impact of the moment the decision about the sender's veracity is made and the sender's facial appearance on police officers' credibility judgments. The Journal of Credibility Assessment and Witness Psychology, 4, 1-36.

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. Psychology, Crime, and Law, 11, 99-122.

Mattson, M., Ryan, D. J., Allen, M., & Miller, V. (2000). Considering organizations as a unique interpersonal context for deception detection: A meta-analytic review. Communication Research Reports, 17, 148-160.

McCornack, S. A., & Levine, T. R. (1990). When lovers become leery: The relationship between suspicion and accuracy in detecting deception. Communication Monographs, 57, 219-230.

Meissner, C. A., & Kassin, S. M. (2002). "He's guilty!": Investigator bias in judgments of truth and deception. Law and Human Behavior, 26, 469-480.

Millar, M. G., & Millar, K. U. (1997). The effects of cognitive capacity and suspicion on truth bias. Communication Research, 24, 556-570.

Mullen, B. (1989). Advanced basic meta-analysis. Hillsdale, New Jersey: Lawrence Erlbaum.

O'Sullivan, M. (2003). The fundamental attribution error in detecting deception: The boy-who-cried-wolf effect. Personality and Social Psychology Bulletin, 29, 1316-1327.

O'Sullivan, M., Ekman, P., & Friesen, W. V. (1988). The effect of comparisons on detecting deceit. Journal of Nonverbal Behavior, 12, 203-215.

Park, H. S., & Levine, T. R. (2001). A probability model of accuracy in deception detection experiments. Communication Monographs, 68, 201-210.

Porter, S., Woodworth, M., & Birt, A. R. (2000). Truth, lies, and videotape: An investigation of the ability of federal parole officers to detect deception. Law and Human Behavior, 24, 643-658.

Rae, G. (1976). Table of A'. Perceptual and Motor Skills, 42, 98.

Reinhard, M. A., Burghardt, K., Sporer, S. L., & Bursch, S. E. (2002). Alltagsvorstellungen über inhaltliche Kennzeichen von Lügen: Selbstberichtete Begründungen bei konkreten Glaubhaftigkeitsurteilen. Zeitschrift für Sozialpsychologie, 33, 169-180.

Ruby, C. L., & Brigham, J. C. (1998). Can Criteria-Based Content Analysis distinguish between true and false statements of African-American speakers? Law and Human Behavior, 22, 369-388.

Santtila, P., Roppola, H., Runtti, M., & Niemi, P. (2000). Assessment of child witness statements using Criteria-Based Content Analysis (CBCA): The effects of age, verbal ability, and interviewer's emotional style. Psychology, Crime, and Law, 6, 159-179.

Schuler, H. (Hrsg.) (2004). Lehrbuch Organisationspsychologie (3. Aufl.). Bern: Huber.

Sporer, S. L. (1996/1998/2004). The Aberdeen Report Judgment Scales (ARJS). Definitions and Answer Sheets. Unpublished Questionnaire. University of Aberdeen, Scotland; University of Giessen, Germany.

Sporer, S. L. (1997a). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. Applied Cognitive Psychology, 11, 373-397.

Sporer, S. L. (1997b). Realitätsüberwachungskriterien und forensische Glaubwürdigkeitskriterien im Vergleich. In L. Greuel, T. Fabian, & M. Stadler (Eds.), Psychologie der Zeugenaussage (pp. 71-85). München: Psychologie Verlags Union.

Sporer, S. L. (1998, March). Detecting deception with the Aberdeen Report Judgment Scales (ARJS): Theoretical development, reliability and validity. Paper presented at the Biennial Meeting of the American Psychology-Law Society, Redondo Beach, CA.

Sporer, S. L. (2004). Reality monitoring and detection of deception. In P. A. Granhag & L. Strömwall (Eds.), Deception detection in forensic contexts (pp. 64-102). Cambridge: University Press.

Sporer, S. L., & Burghardt, S. E. (2004, March). Truth detection with the Aberdeen Report Judgment Scales: The role of planning and rehearsal. Paper presented at the Biennial Meeting of the American Psychology-Law Society in Phoenix, AZ.

Sporer, S. L., & Bursch, S. E. (1996, April). Detection of deception by verbal means: Before and after training. Paper presented at the 38. Tagung experimentell arbeitender Psychologen, Eichstaett, Germany.

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

Sporer, S. L., & Masip, J. (2007). Guidance to detect deception by content cues: Self-efficacy of liars with different types of lies. Final Report to the DAAD. Giessen, Germany, & Salamanca, Spain.

Sporer, S. L., Samweber, M. C., & Stucke, T. S. (2000, March). Twisting the outcome: Discriminating distorted truths from factually experienced events. Paper presented at the Biennial Meeting of the American Psychology-Law Society, New Orleans, LA.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. Applied Cognitive Psychology, 20, 421-446.

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. Psychology, Public Policy, and Law, 13, 1-34.

Sporer, S. L., & Walther, A. (2006, March). Truth detection by content cues: General vs. specific questions. Paper presented at the Meeting of the American Psychology-Law Society in Petersburg, FL.

Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille (Ed.), Credibility Assessment (pp. 135-154). Deventer: Kluwer Academic Publishers.

Steller, M., & Köhnken, G. (1989). Criteria-based statement analysis. In D. C. Raskin (Ed), Psychological methods in criminal investigation and evidence (pp. 217-245). New York: Springer.

Steller, M., Wellershaus, P., & Wolf, T. (1992). Realkennzeichen in Kinderaussagen: Empirische Grundlage der Kriterienorientierten Aussageanalyse. Zeitschrift für Experimentelle und Angewandte Psychologie, 39, 151-170.

Stiff, J. B., Kim, H. J., & Ramesh, C. N. (1992). Truth bias and aroused suspicion in relational deception. Communication Research, 19, 326-345.

Strömwall, L. A., Bengtsson, L., Leander, L., & Granhag, P. A. (2004). Assessing children's statements: The impact of a repeated experience on CBCA and RM ratings. Applied Cognitive Psychology, 18, 653-668.

Toris, D., & DePaulo, B. M. (1984). Effects of actual deception and suspiciousness of deception on interpersonal perceptions. Journal of Personality and Social Psychology, 47, 1063-1073.

Tye, M. C., Amato, S. L., Honts, C. R., Devitt, M. K., & Peters, D. (1999). The willingness of children to lie and the assessment of credibility in an ecologically relevant laboratory setting. Applied Developmental Science, 3, 92-109.

Vrij, A. (2000/2008). Detecting lies and deceit: The psychology of lying and the implications for professional practice. Chichester: John Wiley & Sons.

Vrij, A. (2005). Criteria-based content analysis: A qualitative review of the first 37 studies. Psychology, Public Policy, and Law, 11, 3-41.

Vrij, A., Akehurst, L., Brown, L., & Mann, L. (2006). Detecting lies in young children, adolescents and adults. Legal and Criminological Psychology, 11, 297-312.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004a). Let me inform you how to tell a convincing story: CBCA and reality monitoring scores as a function of age, coaching, and deception. Canadian Journal of Behavioral Science-Revue, 36, 113-126.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004b). Detecting deceit via analyses of verbal and nonverbal behavior in children and adults. Human Communication Research, 30, 8-41.

Vrij, A., Edward, K., & Bull, R. (2001b). Stereotypical verbal and nonverbal responses while deceiving others. Personality and Social Psychology Bulletin, 27, 899-909.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. Journal of Nonverbal Behavior, 24, 239-263.

Vrij, A., & Graham, S. (1997). Individual differences between liars and the ability to detect lies. Expert Evidence, 5, 144-148.

Vrij, A., & Mann, S. (2001a). Telling and detecting lies in a high-stake situation: The case of a convicted murderer. Applied Cognitive Psychology, 15, 187-203.

Vrij, A., & Mann, S. (2001b). Who killed my relative? Police officers' ability to detect real-life high stake lies. Psychology, Crime, and Law, 7, 119-132.

Vrij, A., & Mann, S. (2005). Police use of nonverbal behavior as indicators of deception. In R. E. Riggio & R. S. Feldman (Eds.), Applications of nonverbal communication (pp. 63-94). Mahwah, NJ: Lawrence Erlbaum Associates.

Vrij, A., Mann, S. A., Fisher, R. P., Leal, S., Milne, R., & Bull, R. (2007, published online). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. Law and Human Behavior.

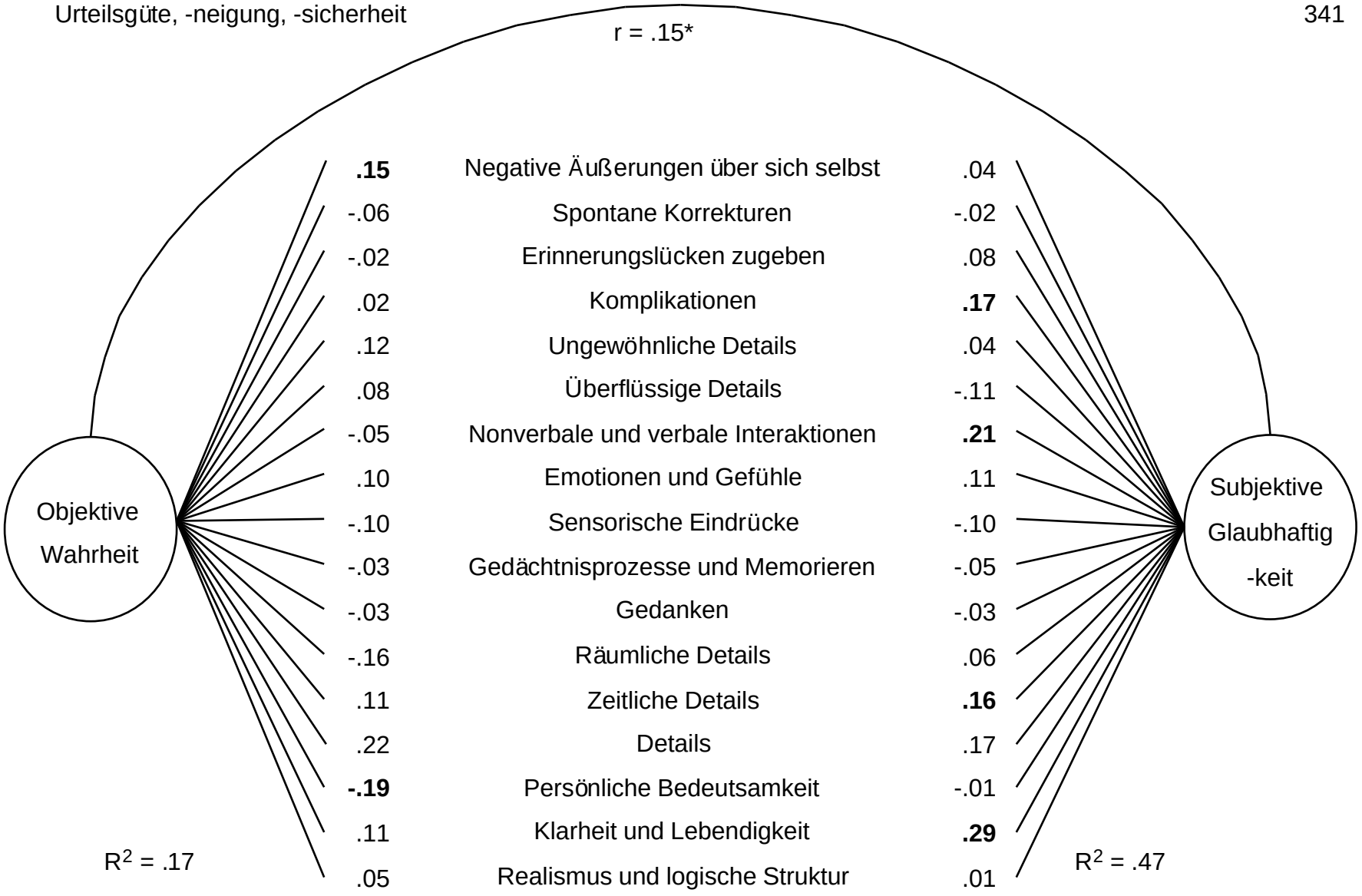
Vrij, A., Mann, S., Kristen, S., & Fisher, R. P. (2007). Cues to deception and ability to detect lies as a function of police interview styles. Law and Human Behavior, 31, 499-518.

Vrij, A., Mann, S., Robbins, E., & Robinson, M. (2006). Police officers ability to detect deception in high stakes situations and in repeated lie detection tests. Applied Cognitive Psychology, 20, 741-755.

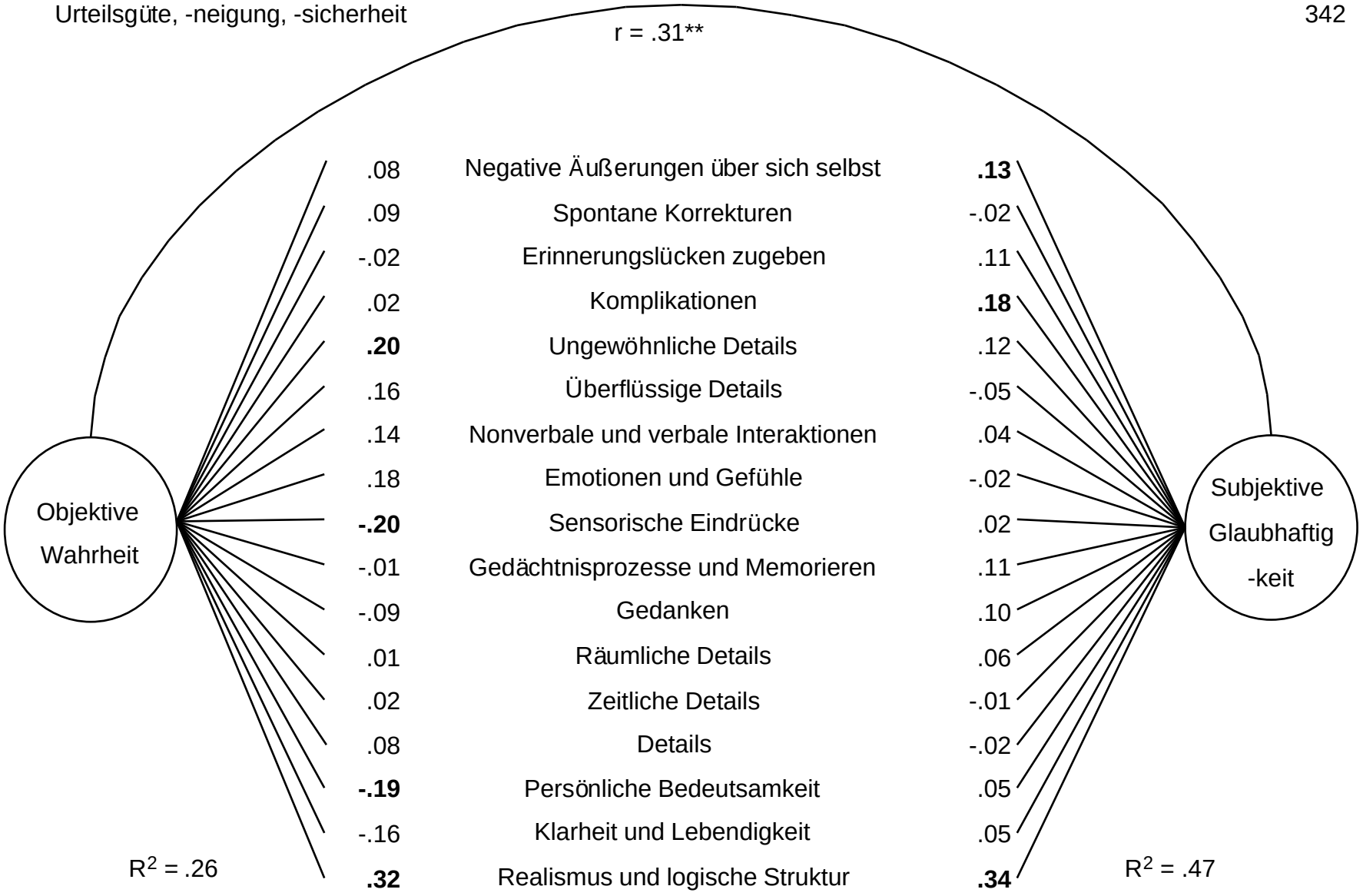
Vrij, A., & Semin, G. R. (1996). Lie experts' beliefs about nonverbal indicators of deception. Journal of Nonverbal Behaviour, 20, 56-80.

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.

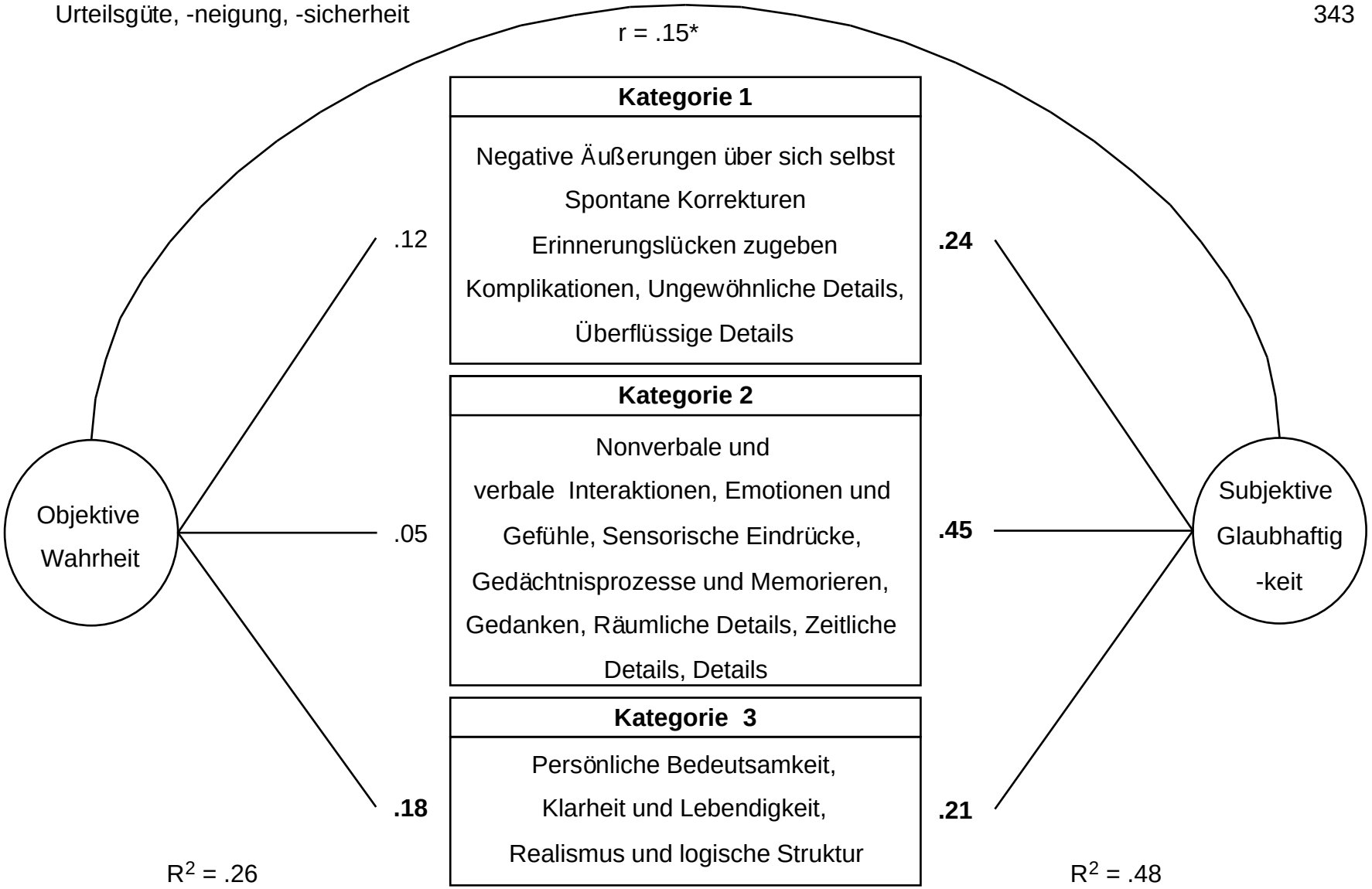
Zuckerman, M., Koestner, R., & Alton, A. O. (1984). Learning to detect deception. Journal of Personality and Social Psychology, 46, 519-528.



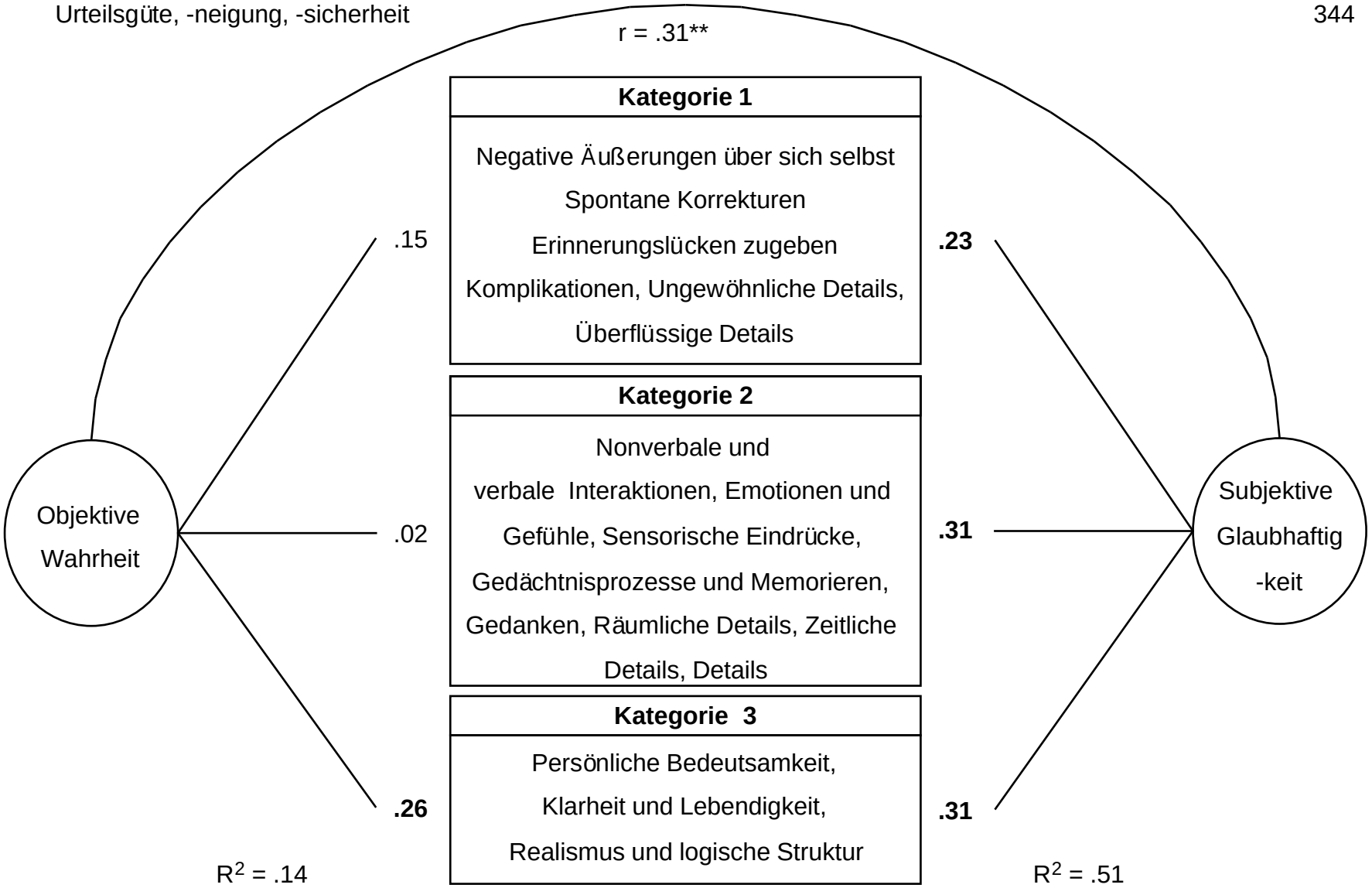
Anhang 5a: Standardisierte beta-Gewichte der multiplen Regressionsanalysen mit den 17 ARJS-STV-G Merkmalen als Prädiktoren für die freien Berichte. Fett gedruckte Koeffizienten unterscheiden sich signifikant von 0.



Anhang 5b: Standardisierte beta-Gewichte der multiplen Regressionsanalysen mit den 17 ARJS-STV-G Merkmalen als Prädiktoren für die Interviews. Fett gedruckte Koeffizienten unterscheiden sich signifikant von 0.



Anhang 5c: Standardisierte beta-Gewichte der multiplen Regressionsanalysen mit den drei Merkmalskategorien als Prädiktoren für die freien Berichte. Fett gedruckte Koeffizienten unterscheiden sich signifikant von 0.



Anhang 5d: Standardisierte beta-Gewichte der multiplen Regressionsanalysen mit den drei Merkmalskategorien als Prädiktoren für die Interviews. Fett gedruckte Koeffizienten unterscheiden sich signifikant von 0.

Anhang 5e

Prozentuale Urteilsrichtigkeiten, Urteilsgüte und -neigung der einzelnen Raterinnen für die freien Berichte

	Ohne Anleitung							Mit Anleitung						
Raterin	Ges	H	CR	G	BR	A'	B''	Ges	H	CR	G	BR	A'	B''
A	52.3	42.9	60.9	40.9	47.7	.54	.35	47.7	81.0	17.4	81.8	47.7	.47	-.91
B	61.4	57.1	65.2	45.5	47.7	.68	.17	56.8	71.4	43.5	63.6	47.7	.64	-.53
C	43.2	56.0	26.3	63.6	56.8	.34	-.56	61.4	38.1	82.6	27.3	47.7	.70	.77
D	54.5	76.0	26.3	75.0	56.8	.53	-.80	50.0	85.7	17.4	84.1	47.7	.55	-.93
E	54.5	52.4	56.5	47.7	47.7	.58	.08	56.8	76.2	39.1	68.2	47.7	.65	-.67
F	43.2	42.9	43.5	50.0	47.7	.38	.01	68.2	95.2	43.5	75.0	47.7	.82	-.93
G	63.6	61.9	65.2	47.7	47.7	.71	.07	59.1	68.0	47.4	61.4	56.8	.64	-.40
H	56.8	90.5	26.1	81.8	47.7	.70	-.93	61.4	72.0	47.4	63.6	56.8	.67	-.48

Anm. n = 44 für jede Raterin. Ges = Prozentuale Urteilsrichtigkeit insgesamt; H = Treffer, CR = Korrekte Zurückweisungen, G = Anteil der als glaubhaft beurteilten an allen Aussagen, BR = Basisrate. Die Raterinnen A&B, C&D, E&F sowie G&H beurteilten jeweils dieselben Aussagen zum selben Zeitpunkt.

Anhang 5f

Prozentuale Urteilsrichtigkeiten, Urteilsgüte und -neigung der einzelnen Raterinnen für die Interviews

	Ohne Anleitung							Mit Anleitung						
Raterin	Ges	H	CR	G	BR	A'	B“	Ges	H	CR	G	BR	A'	B“
A	42.0	45.7	38.1	53.4	52.3	.36	-.15	56.8	95.2	21.7	86.4	47.7	.74	-.97
B	42.0	60.9	21.4	69.3	52.3	.33	-.70	58.0	83.3	34.8	73.9	47.7	.68	-.81
C	59.1	71.7	45.2	63.6	52.3	.65	-.51	61.4	52.4	69.6	40.9	47.7	.68	.35
D	54.5	69.6	38.1	65.9	52.3	.58	-.58	53.4	95.2	15.2	89.8	47.7	.70	-.98
E	48.9	50.0	47.8	51.1	47.7	.48	-.04	59.1	93.5	21.4	86.4	52.3	.71	-.96
F	44.3	45.2	43.5	51.1	47.7	.40	-.04	54.5	95.7	9.5	93.2	52.3	.65	-.99
G	63.6	81.0	47.8	65.9	47.7	.74	-.65	56.8	69.6	42.9	63.6	52.3	.62	-.51
H	55.7	88.1	26.1	80.7	47.7	.68	-.91	60.2	73.9	45.2	64.8	52.3	.67	-.55

Anm. n = 88 für jede Raterin. Ges = Prozentuale Urteilsrichtigkeit insgesamt; H = Treffer, CR = Korrekte Zurückweisungen, G = Anteil der als glaubhaft beurteilten an allen Aussagen, BR = Basisrate. Die Raterinnen A,B,C,D, und E,F,G,H beurteilten jeweils dieselben Aussagen zum selben Zeitpunkt.

DISKUSSION

Die Unterscheidung wahrer und erfundener Aussagen ist in vielen Lebensbereichen von Bedeutung, z.B. in partnerschaftlichen und gesellschaftlichen Beziehungen, in geschäftlichen und politischen Verhandlungen, bei der Auswahl von Bewerbern und bei der Bearbeitung von Anträgen zu Versicherungsleistungen. Sie stellt zudem eine wesentliche Voraussetzung für eine faire Rechtssprechung im Gerichtsverfahren dar. Dies gilt insbesondere für Fälle, bei denen äußere Beweismittel fehlen. Allerdings wurde wiederholt festgestellt, dass Personen kaum dazu in der Lage sind, wahre und erfundene Aussagen zu unterscheiden. Ihre Urteilsgüte liegt nach metaanalytischen Befunden nur geringfügig über dem Zufallsniveau (Bond & DePaulo, 2006; Kraut, 1980). Dies gilt sowohl für Laien, als auch für Berufsgruppen, denen ein Expertenstatus zugeschrieben wird, wie Richter, Polizisten und Psychiater (Aamodt & Custer, 2006). Als Ursache für die schlechte Urteilsgüte wird unter anderem auf fehlerhafte Alltagsvorstellungen zu Täuschungindikatoren verwiesen (z.B. Vrij, 2000/2008).

Ziel der vorliegenden Arbeit war es, solche Alltagsvorstellungen auszudifferenzieren und Möglichkeiten zur Verbesserung der Urteilsgüte aufzuzeigen. Dazu wurden einerseits vermeintliche und andererseits tatsächliche Unterschiede zwischen wahren und erfundenen Aussagen untersucht. In zwei Studien wurden Personen befragt, welche Verhaltensweisen sie subjektiv mit wahren und erfundenen Aussagen assoziieren, während sich zwei weitere Studien auf objektive Täuschungskorrelate konzentrierten. Schließlich wurde in einer fünften Studie überprüft, wie sich empirisch fundiertes Wissen zu inhaltlichen Glaubhaftigkeitsmerkmalen auf die Urteilsgüte auswirkt. Im Folgenden werden die Hauptergebnisse dieser Untersuchungen zusammenfassend diskutiert.

Subjektive Annahmen zu Täuschungskorrelaten

In der ersten Studie wurden Laien (überwiegend Studierende) anhand eines Fragebogens zu ihren subjektiven Täuschungsannahmen befragt. Dieser umfasste insgesamt 73 Items, die auf non- und paraverbales Verhalten sowie auf den Gesamteindruck bezogen waren. Durch die Vorgabe fiktiver Szenarien wurde der Kontext, auf den sich diese Täuschungsannahmen beziehen sollten, experimentell manipuliert.

Die Befunde verwiesen auf starke Diskrepanzen zwischen subjektiven und objektiven Korrelaten von Täuschung. Die befragten Laien erwarteten für 52 der 73 Indikatoren signifikante Unterschiede zwischen Lügnern und wahr aussagenden Personen. Nach den metaanalytischen Befunden von DePaulo, Lindsay, Malone, Muhlenbruck, Charlton und Cooper (2003) können jedoch nur 12 dieser Indikatoren zur Unterscheidung wahrer und erfundener Aussagen beitragen. Dem bisherigen Forschungsstand entsprechend (vgl. Global Deception Research Team, 2006), wurde beispielsweise erwartet, dass Lügner ihren Blick abwenden, obwohl dies objektiv nicht nachzuweisen ist. Selbst in den wenigen Fällen, in denen subjektive und objektive Indikatoren der Richtung nach korrespondierten, wurde das Ausmaß der tatsächlichen Verhaltensunterschiede, d.h. ihre relative Zu- und Abnahme, subjektiv überschätzt.

Metaanalytisch wurden verschiedene Moderatoren hinsichtlich der objektiven Differenzierungskraft non- und paraverbaler Indikatoren nachgewiesen (DePaulo et al., 2003; Sporer & Schwandt, 2006, 2007). Bei der Entwicklung der fiktiven Szenarien wurden zwei dieser Moderatoren aufgegriffen. Die Motivation erfolgreich zu täuschen wurde durch drei verschiedene Täuschungssituationen variiert. Dabei wurde ausgeführt, dass eine Person lügt, um sich wegen einer Verspätung zu rechtfertigen (Alltagssituation), um eine ausereheliche Affäre zu verheimlichen (Affärensituation) oder um die Verantwortung für ein Totschlagdelikt von sich zu weisen (Verbrechenssituation). Die Motivation zu überzeugen wurde für die Affären- und Verbrechenssituation als gleichermaßen hoch, doch für die

Alltagssituation signifikant geringer eingestuft. Dennoch variierten die subjektiven Täuschungsannahmen kaum. So wurden nur 10 der 73 Indikatoren in Abhängigkeit von der vorgegebenen Täuschungssituation unterschiedlich bewertet. Dies war vor allem auf Unterschiede in der Stärke der angenommenen Effekte zurückzuführen. Die Befunde weisen darauf hin, dass ernste Lügen mit mehr kognitiver Anstrengung und Stress assoziiert werden als Alltagslügen (vgl. Vrij & Taylor, 2003).

Zudem wurde durch die fiktiven Szenarien das Ausmaß der vermeintlichen Vorbereitungszeit des Lügners erfolgreich manipuliert. In Abhängigkeit von der Gelegenheit zur Vorbereitung wurden neun Indikatoren unterschiedlich bewertet. Die meisten Unterschiede zeigten sich dabei im sprachlichen Bereich, einige wenige auch hinsichtlich des Gesamteindrucks. Beispielsweise war die subjektive Annahme, dass Lügner einen nervösen Gesamteindruck hinterlassen, in Bedingungen ohne Vorbereitung stärker ausgeprägt als mit Vorbereitung. Allerdings wurden Verhaltensweisen im Körperbereich, die oftmals mit Nervosität assoziiert werden, nicht unterschiedlich bewertet.

Insgesamt wurden kontextuelle Determinanten von Aussagen bei den subjektiven Annahmen kaum berücksichtigt. Dies sollte bei Trainings zur Entdeckung von Täuschung aufgegriffen werden. Nach den vorliegenden Untersuchungsbefunden wäre vor allem darauf hinzuweisen, dass allgemein nur geringe Verhaltensunterschiede im non- und paraverbalen Bereich zu erwarten sind. Dies gilt insbesondere, wenn die lügende Person nur wenig motiviert ist zu überzeugen, oder auch Gelegenheit hatte sich gut vorzubereiten.

In der zweiten Studie wurde das gleiche Paradigma verwendet, um subjektive Annahmen zu inhaltlichen Glaubhaftigkeitsmerkmalen zu erfassen. Der Fragebogen umfasste 52 Merkmale, die aus den Aberdeen Report Judgment Scales (ARJS; Sporer, 1996/1998/2004) abgeleitet worden waren. Erneut wurden das Ausmaß der Vorbereitungszeit und die Täuschungssituation über die Vorgabe fiktiver Szenarien variiert. Dabei wurden die bereits in der ersten Studie

verwendeten Alltags-, Affären- oder Verbrechenszenarien vorgegeben. Zusätzlich wurde ein Szenario konzipiert, bei dem eine Frau nach gründlicher Vorbereitung gegenüber der Polizei fälschlicherweise aussagt vergewaltigt worden zu sein. Dadurch sollte überprüft werden, ob es spezifische subjektive Annahmen für Sexualdelikte gibt (vgl. Niehaus, Krause & Schmidke, 2005). Eine weitere Gruppe von Laien beantwortete den Fragebogen ohne vorgegebene Informationen zum situativen Kontext und diente als Kontrollgruppe.

Die meisten inhaltlichen Aussagemerkmale wurden gemäß ihrer objektiven Differenzierungskraft mit wahren Aussagen assoziiert. Dies galt beispielsweise für die ARJS-Skalen Realismus und logische Struktur, nonverbale und verbale Interaktionen, Details und Gedanken. Hingegen wurden Komplikationen und ungewöhnliche Details dem bisherigen Forschungsstand entsprechend oftmals mit erfundenen Aussagen assoziiert. Empirische Untersuchungen zu den ARJS zeigten jedoch, dass diese Merkmale objektiv als Wahrheitsindikatoren aufzufassen sind (Sporer, 1998; Sporer & Burghardt, 2004; Sporer & Walther, 2006). Aufgrund der gegensätzlichen Alltagsvorstellungen sind für diese Skala vermutlich keine täuschungsstrategischen Beeinträchtigungen ihrer Validität zu befürchten. Die Validität einiger Glaubhaftigkeitsmerkmale wurde sogar darauf zurückgeführt, dass Personen sie mit Täuschung assoziieren und daher gezielt vermeiden, wenn sie falsch aussagen. Dies gilt sowohl für die stereotypkonträren Merkmale der Criteria-Based Content Analysis (CBCA) als auch für die ARJS-Skala Fehler und sozial Unerwünschtes. Die subjektiven Annahmen für diese Skala variierten in Abhängigkeit von der Untersuchungssituation. Während Fehler und sozial Unerwünschtes für die Verspätungssituation tendenziell mit erfundenen Aussagen assoziiert wurden, wurde die Skala in der Bedingung zum Sexualdelikt mit wahren Aussagen assoziiert. Demnach wird die Argumentation, dass solche Merkmale subjektiv als Lügenindikatoren gewertet werden (DePaulo et al., 2003; Sporer, 1996/1998/2004; Steller & Köhnken, 1989), durch die Ergebnisse der vorliegenden Arbeit nicht unterstützt.

Die subjektiven Annahmen unterschieden sich in Abhängigkeit von der vorgegebenen Täuschungssituation. Insbesondere für das Szenario zum Sexualdelikt waren spezifische subjektive Annahmen festzustellen (vgl. Niehaus et al., 2005). Allerdings wurde dabei die objektive Differenzierungskraft der inhaltlichen Aussagemerkmale meist überschätzt. Dies erscheint vor allem deswegen problematisch, weil das Szenario darauf verwies, dass die lügende Person lediglich die Berichte einer anderen Person nachzuerzählen brauchte. Die befragten Laien beachteten offenbar nicht, dass dies einfacher sein sollte, als eine Falschaussage frei zu erfinden. Zudem unterschieden sich die subjektiven Annahmen nicht in Abhängigkeit von der Gelegenheit zur Vorbereitung (s. auch Taylor & Vrij, 2000). Personen scheinen demnach unterschiedliche Schwierigkeitsgrade beim Lügen nicht zu berücksichtigen. Dieser Befund steht im Kontrast zu den freien Angaben von Laien, nach denen es durchaus als relevant erachtet wird, ob eine Aussage spontan vorgebracht oder einstudiert wirkt (Granhag & Strömwall, 2000).

Schließlich wurde überprüft, ob Personen ohne konkrete Situationsvorgabe an Täuschungssituationen hoher Relevanz denken (Lakhani & Taylor, 2003; Zuckerman, Koestner & Driver, 1981). Die subjektiven Annahmen der Kontrollgruppe unterschieden sich jedoch vor allem von den Untersuchungsgruppen, denen das Szenario eines Totschlags- oder Sexualdelikts vorgegeben wurde. Demnach schienen Studierende, die ihre subjektiven Annahmen ohne Situationsvorgabe berichteten, zumindest nicht an strafrechtlich relevante Täuschungssituationen zu denken. Andererseits wäre dies durchaus bei Personen denkbar, die solchen Situationen beruflich ausgesetzt sind. Diese Frage lässt sich jedoch nicht klären, wenn die subjektiven Annahmen verschiedener Berufsgruppen ohne Situationsvorgabe erfasst werden (Akehurst, Köhnken, Vrij & Bull, 1996; Strömwall & Granhag, 2003).

Insgesamt scheinen vor allem die Alltagsvorstellungen zu non- und paraverbalen Verhaltensweisen fehlerhaft zu sein. Vrij und Granhag (2007) stellten

fest, dass auch Manuale für polizeiliche Vernehmungen oftmals empfehlen, auf Verhaltensweisen zu achten, die nicht valide zwischen wahren und erfundenen Aussagen unterscheiden. Daher sollte im Rahmen von Trainings die subjektive Annahme revidiert werden, dass starke Unterschiede zwischen wahr aussagenden und lügenden Personen zu erwarten sind. Hingegen wurde für die inhaltlichen Aussagemerkmale eine stärkere Korrespondenz zwischen den subjektiven Annahmen und ihrer objektiven Differenzierungskraft festgestellt. Dies könnte darauf hinweisen, dass Personen ihr verbales Verhalten bewusster wahrnehmen und stärker reflektieren als ihr non- und paraverbales. Dennoch erlauben es inhaltliche Aussagemerkmale besser als non- und paraverbale Verhaltensweisen zwischen wahren und erfundenen Aussagen zu differenzieren. Die vergleichsweise realistischen Alltagsvorstellungen nivellieren offenbar nicht die Validität der inhaltlichen Aussagemerkmale.

Zudem stehen die vorliegenden Untersuchungsbefunde im Einklang mit der Argumentation, dass für Sexualdelikte spezifische subjektive Täuschungsannahmen vorliegen. Insgesamt schien der Kontext einer Aussage jedoch nur wenig Beachtung zu finden. Dies sollte bei Trainings zur Entdeckung von Täuschung aufgegriffen werden, um Diskrepanzen zwischen subjektiven und objektiven Täuschungsindikatoren abzubauen. Schließlich wäre es ein interessantes Forschungsanliegen, die subjektiven Annahmen zu Täuschungskorrelaten personengebunden zu untersuchen. Möglicherweise berücksichtigen Personen die kontextuellen Besonderheiten von Aussagen eher, wenn sie das Verhalten von ihnen vertrauten Personen einschätzen sollen. Schließlich sollten subjektive Alltagsvorstellungen auch vor dem Hintergrund weiterer Moderatoren der objektiven Differenzierungskraft systematisch ausdifferenziert werden.

Objektive inhaltliche Indikatoren von Täuschung

Zudem wurde untersucht, ob sich inhaltliche Glaubhaftigkeitsmerkmale reliabel erfassen lassen und eine valide Unterscheidung wahrer und erfundener Aussagen erlauben. Als Stimulusmaterial wurden Transkripte von 176 freien Berichten und 176 Interviews einer Studie von Sporer und Burghardt (2004) verwendet. Diese wurden entweder unvorbereitet oder nach einer kurzen Vorbereitungszeit abgegeben und bezogen sich auf persönlich bedeutsame Lebensereignisse mit unterschiedlicher Valenz. Die inhaltliche Qualität der Aussagen wurde in der dritten Studie von vier Beurteilerinnen anhand der ARJS, in der vierten von acht Beurteilerinnen anhand einer Kurzform derselben (Aberdeen Report Judgment Scales--Short Training Version--German, ARJS-STV-G) erfasst.

Inter-Rater-Reliabilität

Für die meisten ARJS-Skalen wurden gute Inter-Rater-Reliabilitäten erzielt. Dabei sind die guten Befunde hinsichtlich der ARJS-Skala Gedanken hervorzuheben. Auf der Grundlage des Realitätsüberwachungsansatzes (Johnson & Raye, 1981) wurde argumentiert, dass kognitive Operationen eher bei erfundenen als bei wahren Aussagen vorzufinden sein sollten. Allerdings erwies es sich als schwierig dieses Merkmal reliabel zu erfassen, und die Befunde zu dessen Validität sind äußerst widersprüchlich (Sporer, 1997a, 1997b; Sporer & Küpper, 1995; Vrij, Akehurst, Soukara & Bull, 2004; vgl. Masip, Sporer, Garrido & Herrero, 2005; Sporer, 2004). Im Gegensatz dazu ist die ARJS-Skala Gedanken als Wahrheitsindikator konzipiert, präzise operationalisiert, und wird getrennt von der ARJS-Skala Memorieren und Gedächtnis erfasst. Nach den vorliegenden Befunden läßt sich dadurch eine hohe Inter-Rater-Reliabilität gewährleisten (Barnier, Sharman, McKay, & Sporer, 2005; Sporer, 1998, Sporer, Bursch, Schreiber, Weiss, Höfer, Sievers & Köhnken, 2000; Sporer & Walther, 2006). Zudem wurde demonstriert, dass sich die Inter-Rater-Reliabilität verbessert, wenn mehrere unabhängige Beurteilungen zusammengefasst werden. Für

individualdiagnostische Anwendungen wären daher drei bis vier unabhängige Rater zu empfehlen.

Für die ARJS-STV-G fielen die Inter-Rater-Reliabilitäten vergleichsweise geringer aus. Dies ist vermutlich darauf zurückzuführen, dass die Beurteilerinnen nur eine kurze Anleitung zu den Glaubhaftigkeitsmerkmalen erhielten. Die Zusammenfassung von vier unabhängigen Beurteilungen scheint jedoch eine zufrieden stellend reliable Erfassung der ARJS-STV-G-Merkmale zu gewährleisten.

Validität

Für neun der 13 ARJS-Skalen waren erwartungsgemäße Unterschiede zwischen wahren und erfundenen Aussagen nachweisbar. Diese Skalen wurden auf der Grundlage verschiedener Theorien und Forschungsbefunde konzipiert. Beispielsweise wurde die Validität der Skala Klarheit und Lebendigkeit auf Forschungsbefunde zum autobiographischen Gedächtnis und zum Realitätsüberwachungsansatz zurückgeführt (vgl. Brewer, 1986, 1988, 1996, sowie Johnson & Raye, 1981; Larsen, 1998; McGinnis & Roberts, 1996; Suengas & Johnson, 1988). Allerdings scheinen auch Modifikationen notwendig zu sein, um den Realitätsüberwachungsansatz auf die interpersonelle Ebene zu übertragen. So war beispielsweise die Skala Gedanken erwartungsgemäß bei wahren Aussagen stärker ausgeprägt als bei erfundenen. Dies widerspricht den Vorhersagen des Realitätsüberwachungsansatzes, dass kognitive Operationen als Lügenindikatoren zu werten seien. Die ARJS-Skala Gedanken soll jedoch Merkmale umfassen, die auf stützende Erinnerungen hinweisen (vgl. Johnson, 1985) und daher die Hypothese eines persönlichen Erfahrungsbezugs stärken. Dass Fehler und Sozial Unerwünschtes subjektiv mit Täuschung assoziiert werden (Sporer, 1996/1998/2004), ließ sich im Rahmen der Studie 2 nicht belegen. Dennoch erwies sich die ARJS-Skala Fehler und Sozial Unerwünschtes als valide. Es erscheint plausibel, dass sozialpsychologische Überlegungen den gedächtnispsychologischen Ansatz bei der Konzeption von Glaubhaftigkeitsmerkmalen sinnvoll ergänzen können. Eine empirische

Untermauerung dieser Argumentation steht jedoch noch aus. Schließlich trugen auch ARJS-Skalen, die den CBCA- und den Realitätsüberwachungsmerkmalen konzeptionell ähnlich sind, zur Unterscheidung wahrer und erfundener Aussagen bei.

Von den 17 ARJS-Merkmalen der Kurzform differenzierten 10 signifikant zwischen wahren und erfundenen Aussagen. Die Effektstärken fielen dabei kaum geringer aus als für die Langform. Allerdings war für die freien Berichte kein Haupteffekt des Wahrheitsstatus auf die ARJS-STV-G-Merkmale mehr nachzuweisen, wenn die Wortanzahl kontrolliert wurde. Die höhere Qualität wahrer im Vergleich zu erfundenen freien Berichten war demnach vor allem auf den erhöhten Umfang der Aussagen zurückzuführen. Das Ziel der Aussageanalyse, wahre und erfundene Aussagen zu unterscheiden, wurde dadurch jedoch nicht beeinträchtigt. Schließlich wurde für das verwendete Stimulusmaterial festgestellt, dass wahre Aussagen umfangreicher sind als erfundene (vgl. Strömwall, Bengtsson, Leander & Granhag, 2004; Vrij, Edward, Roberts & Bull, 2000).

Moderatoren der Validität

Die Aussagequalität wurde bei vorbereiteten freien Berichten höher eingeschätzt als bei unvorbereiteten. Dieser Effekt ließ sich für die ARJS-Merkmale multivariat absichern und war univariat auf vier Merkmale zurückzuführen. Auch für acht ARJS-STV-G-Merkmale zeigten sich entsprechende Mittelwertsunterschiede zwischen vorbereiteten und unvorbereiteten Aussagen. Multivariat wurde jedoch kein signifikanter Effekt der Vorbereitung auf die Aussagequalität festgestellt. Für die Interviews, die eine Woche später erhoben wurden, waren keine vorbereitungsbedingten Unterschiede mehr festzustellen. Offenbar ist es weniger entscheidend ob, sondern vor allem wann eine Aussage vorbereitet wurde. Wechselwirkungseffekte zwischen dem Wahrheitsstatus und der Vorbereitungszeit waren jedoch nicht festzustellen. Demnach wären keine vorbereitungsbedingten Beeinträchtigungen der Validität zu befürchten. Allerdings

ist der Forschungsstand hierzu insgesamt als widersprüchlich zu bewerten (vgl. Alonso-Quecuty, 1992; Sporer, 1997a, 1998; Sporer & Burghardt, 2004).

Auch die Valenz der geschilderten Ereignisse hatte einen Effekt auf die Aussagequalität. Negative Ereignisse erzielten höhere ARJS-Beurteilungen und zumindest für die Interviews auch höhere ARJS-STV-G-Beurteilungen als positive. Die Valenz der geschilderten Ereignisse moderierte jedoch nicht den Zusammenhang zwischen dem Wahrheitsstatus und der Aussagequalität. Demnach scheinen die inhaltlichen Aussagemerkmale auch auf positive Ereignisse anwendbar zu sein. Dies unterstützt die Zielsetzung eines breiten Geltungsbereichs der ARJS (vgl. Sporer, 1996/1998/2004), der für die CBCA-Merkmale in Frage gestellt wurde (Landry & Brigham, 1992; Steller, 1989).

Zusammenfassend konnten die inhaltlichen Glaubhaftigkeitsmerkmale valide zwischen wahren und erfundenen Aussagen differenzieren. Dies galt für eine Vielzahl persönlich bedeutsamer Lebensereignisse, so dass ein breiter Anwendungsbereich denkbar ist. Eine Replikation dieser Befunde bei der Untersuchung kindlicher Aussagen steht jedoch aus.

Urteilsgüte

Bislang blieb ungeklärt, ob das Wissen um die empirisch fundierten ARJS-STV-G-Merkmale eine Verbesserung der Urteilsgüte bewirkt (Sporer et al., 2000). Daher wurden in der fünften Studie die subjektiven Glaubhaftigkeitsurteile naiver und ARJS-STV-G-angeleiteter Beurteilerinnen vergleichend betrachtet. So schätzten die acht Beurteilerinnen die Glaubhaftigkeit für einen Teil des Stimulusmaterials naiv und später für einen anderen Teil ARJS-STV-G-angeleitet ein.

Für die naiven Beurteilungen ergab sich eine prozentuale Urteilsrichtigkeit von 50.6% für die freien Berichte und von 51.1% für die Interviews. Die Urteilsgüte lag damit auf Zufallsniveau. Dieser Befund steht im Einklang mit einer Vielzahl an Untersuchungen, die demonstrierten, dass Laien kaum dazu in der Lage sind

wahre und erfundene Aussagen zu unterscheiden (vgl. Bond & DePaulo, 2006; Kraut, 1980). Zudem kann aufgrund der zufälligen Urteilsgüte ausgeschlossen werden, dass bei dem verwendeten Stimulusmaterial Unterschiede zwischen wahren und erfundenen Aussagen allzu offensichtlich waren. Für die ARJS-STV-G-angeleiteten Beurteilungen der Interviews wurde eine überzufällige Urteilsgüte von 61.9% nachgewiesen. Das Wissen um die Glaubhaftigkeitsmerkmale bewirkte offenbar eine Verbesserung in der Einschätzung wahrer Aussagen. Während die Urteilsgüte bei erfundenen Aussagen mit nur 29.5% gering blieb, zeigte sich bei wahren Aussagen eine beeindruckende prozentuale Urteilsrichtigkeit von 94.3%. Diese war trotz einer verstärkten Tendenz, Aussagen als glaubhaft zu bewerten, als überzufällig zu bewerten. Für die freien Berichte fiel die prozentuale Urteilsrichtigkeit bei den ARJS-STV-G-angeleiteten Beurteilungen mit 56.8% nur augenscheinlich höher aus als bei den naiven. Dies ist vermutlich darauf zurückzuführen, dass die Glaubhaftigkeitsmerkmale bei den Interviews häufiger vorzufinden waren als bei den freien Berichten.

Zudem gab es bei den freien Berichten vorbereitungsbedingte Unterschiede in der Entscheidungsrichtigkeit für wahre und erfundene Aussagen ohne und mit der ARJS-STV-G-Anleitung. Im Gegensatz zu den naiven Beurteilungen verschlechterte sich durch die Vorbereitung die ARJS-STV-G-angeleitete Einschätzung der erfundenen Aussagen, während sich die der wahren verbesserte. Dies ist vermutlich darauf zurückzuführen, dass es vorbereiteten Stimuluspersonen gelang, die Qualität sowohl für wahre als auch für erfundene Aussagen zu erhöhen. Offenbar wird durch Vorbereitung nicht nur die Fähigkeit Geschichten zu erfinden, sondern auch der Abruf autobiographischer Erinnerungen gefördert. Ob vorbereitungsbedingte Anreicherungen der Qualität als positiv oder negativ zu erachten sind, hängt auch von der relativen Wichtigkeit der Urteilsgüte für wahre und erfundene Aussagen ab.

Nach den vorliegenden Befunden können die ARJS-STV-G bei der richtigen Einschätzung wahrer Aussagen hilfreich sein. Die Glaubhaftigkeitsmerkmale

lassen sich mit wenig Aufwand vermitteln und einfach anwenden. Daher könnte die ARJS-STV-G eine interessante Weiterbildungsmaßnahme für verschiedene Berufsgruppen darstellen. Da es zur Verstärkung liberaler Urteilstendenzen kommen kann, wäre die ARJS-STV-G-Anleitung insbesondere Polizisten zu empfehlen, die zu einem Lügenbias neigen (vgl. Garrido, Masip & Herrero, 2004). Die höchste Urteilsgüte erzielten jedoch Experten, die eine vollständige ARJS-Analyse durchführten. Allerdings ist mit einer vollständigen ARJS-Analyse ein erheblicher Schulungs- und Durchführungsaufwand verbunden. Daher kommen in erster Linie psychologische Sachverständige, die sich auf Glaubhaftigkeitsbegutachtungen spezialisiert haben, als potentielle Anwender in Frage. Es ist als besonderer Vorteil der ARJS zu werten, dass sowohl eine Langform als auch eine ökonomische Kurzform vorliegt. Schließlich sind in unserem Rechtssystem viele Personen an der Wahrheitsfindung beteiligt, die über unterschiedliche zeitliche Kapazitäten für psychologische Weiterbildungsmaßnahmen verfügen.

Es bleibt festzuhalten, dass die Wahrheitsfindung eine der wichtigsten Aufgaben, aber auch schwierigsten Herausforderungen in unserem Rechtssystem darstellt. Die vorliegende Arbeit differenzierte Alltagsvorstellungen zu Täuschung aus, überprüfte die Validität theoretisch fundierter Aussagemerkmale und analysierte die Effekte einer Anleitung zu diesen Merkmalen. Die Befunde liefern Anregungen für die Gestaltung von Trainings zur Entdeckung von Täuschung, werfen jedoch auch weiterführende Forschungsfragen auf. Die ARJS-Merkmale weiterhin zu untersuchen stellt sicherlich einen vielversprechenden Forschungsansatz dar.

Literatur

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar: A meta-analysis of individual differences in detecting deception. Forensic Examiner, 15, 6-11.

Akehurst, L., Köhnken, G., Vrij, A., & Bull, R. (1996). Lay persons' and police officers' beliefs regarding deceptive behaviour. Applied Cognitive Psychology, 10, 461-471.

Alonso-Quecuty, M. (1992). Deception detection and reality monitoring: A new anser to an old question? In F. Loesel, D. Bender, & T. Bliesener (Eds.), Psychology and Law: International perspectives (pp. 228-332). Berlin: Walter de Gruyter.

Barnier, A. J., Sharman, S. J., McKay, L., & Sporer, S. L. (2005). Discriminating adults' genuine, imagined, and deceptive accounts of positive and negative childhood events. Applied Cognitive Psychology, 19, 985-1001.

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. Personality and Social Psychology Review, 10, 214-234.

Brewer, W. F. (1986). What is autobiographical memory? In D. C. Rubin (Ed.), Autobiographical memory (pp. 25-49). Cambridge: Cambridge University Press.

Brewer, W. F. (1988). Memory for randomly sampled autobiographical events. In U. Neisser & E. Winograd (Eds.), Remembering reconsidered: Ecological and traditional approaches to the study of memory (pp. 21-90). Cambridge: Cambridge University Press.

Brewer, W. F. (1996). What is recollective memory? In D. C. Rubin (Ed.), Remembering our past: Studies in autbiographical memory (pp. 19-66). New York: Cambridge University Press.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. Psychological Bulletin, 129, 74-112.

Garrido, E., Masip, J., & Herrero, C. (2004). Police officers' credibility judgments: Accuracy and estimated ability. International Journal of Psychology, 39, 254-275.

Global Deception Research Team (2006). A world of lies. Journal of Cross-Cultural Psychology, 37, 60-74.

Granhag, P. A., & Strömwall, L. A. (2000). Effects of preconceptions on deception detection and new answers to why lie-catchers often fail. Psychology, Crime, and Law, 6, 197-218.

Johnson, M. K. (1985). The origin of memories. In P. C. Kendall (Ed.), Advances in cognitive-behavioral research and therapy (Vol. 4, pp. 1-27). New York: Academic Press.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. Psychological Review, 88, 67-85.

Kraut, R. (1980). Humans as lie detectors: Some second thoughts. Journal of Communication, 30, 209-216.

Lakhani, M., & Taylor, R. (2003). Beliefs about the cues to deception in high- and low-stake situations. Psychology, Crime, and Law, 9, 357-368.

Landry, K., & Brigham, J. C. (1992). The effect of training in criteria-based content analysis on the ability to detect deception in adults. Law and Human Behavior, 16, 663-675.

Larsen, S. F. (1998). What is it like to remember? On phenomenal qualities of memory. In C. P. Thomson, D. J. Herrmann, D. Bruce, J. D. Read, D. G. Payne, & M. P. Toglia (Eds.), Autobiographical memory: Theoretical and applied perspectives (pp. 163-190). Mahwah, New Jersey: Lawrence Erlbaum.

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. Psychology, Crime, and Law, 11, 99-122.

McGinnis, D., & Roberts, P. (1996). Qualitative characteristics of vivid memories attributed to real and imagined experiences. American Journal of Psychology, 109, 59-77.

Niehaus, S., Krause, A., & Schmidke, J. (2005). Täuschungsstrategien bei der Schilderung von Sexualstraftaten. Zeitschrift für Sozialpsychologie, 36, 175-187.

Sporer, S. L. (1996/1998/2004). The Aberdeen Report Judgment Scales (ARJS). Definitions and answer sheets. Unpublished Questionnaire. University of Aberdeen, Scotland; University of Giessen, Germany.

Sporer, S. L. (1997a). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experiences events. Applied Cognitive Psychology, 11, 373-397.

Sporer, S. L. (1997b). Realitätsüberwachungskriterien und forensische Glaubwürdigkeitskriterien im Vergleich. In L. Greuel, T. Fabian, & M. Stadler (Eds.), Psychologie der Zeugenaussage (S. 71-85). München: Psychologie Verlags Union.

Sporer, S. L. (1998, March). Detecting deception with the Aberdeen Report Judgment Scales (ARJS): Theoretical development reliability and validity. Paper presented at the Biennial Meeting of the American Psychology-Law Society, Redondo Beach, CA.

Sporer, S. L. (2004). Reality monitoring and detection of deception. In P. A. Granhag & L. Strömwall (Eds.), Deception detection in forensic contexts (pp. 64-102). Cambridge: University Press.

Sporer, S. L., & Burghardt, K. (2004, March). Truth detection with the Aberdeen Report Judgment Scales: The role of planning and rehearsal. Paper presented at the Biennial Meeting of the American Psychology-Law Society, Phoenix, AZ.

Sporer, S. L., Bursch, S. E., Schreiber, N., Weiss, P. E., Höfer, E., Sievers, K., & Köhnken, G. (2000). Detecting deception with the Aberdeen Report Judgment Scales (ARJS): Inter-Rater Reliability. In A. Czerederecka, T. Jaskiewicz-Obydzinska, & J. Wojcikiewicz (Eds.), Forensic psychology and law (pp. 197-204). Krakow: Institute of Forensic Research Publishers.

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

Sporer, S. L., & Masip, J. (2007). Guidance to detect deception by content cues: Self-efficacy of liars with different types of lies. Final Report to the DAAD. Giessen, Germany, & Salamanca, Spain.

Sporer, S. L., & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. Psychology, Public Policy, and Law, 13, 1-34.

Sporer, S. L., & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. Applied Cognitive Psychology, 20, 421-446.

Sporer, S. L., & Walther, A. (2006, March). Truth detection by content cues: General vs. specific questions. Paper presented at the Meeting of the American Psychology-Law Society in Petersburg, FL.

Steller, M. (1989). Recent developments in statement analysis. In J. C. Yuille (Ed.), Credibility Assessment (pp. 135-154). Deventer: Kluwer Academic Publishers.

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

Strömwall, L. A., Bengtsson, L., Leander, L., & Granhag, P. A. (2004). Assessing children's statements: The impact of a repeated experience on CBCA and RM ratings. Applied Cognitive Psychology, 18, 653-668.

Strömwall, L. A., & Granhag, P. A. (2003). How to detect deception? Arresting the beliefs of police officers, prosecutors and judges. Psychology, Crime, and Law, 9, 19-36.

Suengas, A. G., & Johnson, M. K. (1988). Qualitative effects of rehearsal on memories for perceived and imagined complex events. Journal of Experimental Psychology: General, 117, 377-389.

Taylor, R., & Vrij, A. (2000). The effects of varying stake and cognitive complexity on beliefs about the cues to deception. International Journal of Police Science and Management, 3, 111-123.

Vrij, A. (2000/2008). Detecting lies and deceit: The psychology of lying and the implications for professional practice. Chichester: John Wiley & Sons.

Vrij, A., Akehurst, L., Soukara, S., & Bull, R. (2004). Let me inform you how to tell a convincing story: CBCA and reality monitoring scores as a function of age, coaching, and deception. Canadian Journal of Behavioral Science-Revue, 36, 113-126.

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. Journal of Nonverbal Behavior, 24, 239-263.

Vrij, A., & Granhag, P. A. (2007). Interviewing to detect deception. In S. A. Christianson (Ed.), Offenders' memories of violent crimes (pp. 279-304). Chicheser, England: John Wiley & Sons, Ltd.

Vrij, A., & Taylor, R. (2003). Police officers' and students' beliefs about telling and detecting trivial and serious lies. International Journal of Police Science and Management, 5, 1-9.

Zuckerman, M., Koestner, R., & Driver, R. (1981). Beliefs about cues associated with deception. Journal of Nonverbal Behavior, 6, 105-114.

ERKLÄRUNG

Ich erkläre: Ich habe die vorgelegte Dissertation selbständig und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht.

Gießen, im Oktober 2008

Maike Miriam Breuer