Die psychometrische Qualität von Realkennzeichen

Peter Steck, Max Hermanutz, Bianca Lafrenz, Domenica Schwind, Stefanie Hettler, Barbara Maier, Susanne Geiger.

✉
Prof. Dr. Peter Steck, Fachbereich Psychologie, Universität Konstanz, Fach 31, 78457 Konstanz
Prof. Dr. Max Hermanutz, Hochschule für Polizei, Villingen-Schwenningen
Dipl.-Psych. Bianca Lafrenz, Max-Planck-Institut für ausländisches und internationales Strafrecht, Freiburg
Dipl.-Psych. Domenica Schwind, Gesellschaft für wissenschaftliche Gerichts- und Rechtspsychologie, München
Dipl.-Psych. Stefanie Hettler, Fachbereich Psychologie, Universität Konstanz
Dipl.-Psych. Barbara Maier, Fachbereich Psychologie, Universität Konstanz
Dipl.-Psych. Susanne Geiger, Fachbereich Psychologie, Universität Konstanz

Kurztitel: Psychometrie und Realkennzeichen

Zusammenfassung

Die vorliegende Arbeit fasst zwei Untersuchungen zu psychometrischen Eigenschaften der in der aussagepsychologischen Diagnostik verwendeten merkmalsorientierte Inhaltsanalyse zusammen. Die erste Untersuchung war Teil einer größeren Laborstudie zur Validität inhaltsanalytischer Beurteilungen der Wahrhaftigkeit von Zeugenaussagen mit 60 Versuchspersonen, die verschiedenen Bedingungen von Wahr- und Falschaussagen zugeordnet waren, die zweite, eine Felduntersuchung, erfasste 138 Gutachten eines gerichtspsychologischen Institutes. In beiden Studien wurden die Items der merkmalsorientierten Inhaltsanalyse, die so genannten Realkennzeichen, einer Itemanalyse nach dem Modell der klassischen Testtheorie mit Reliabilitätsschätzung für den Score aggregierter Realkennzeichen unterzogen. Die Ergebnisse differierten beträchtlich: Während in der Laborstudie nur ein Teil der für die Realkennzeichen errechneten Trennschärfekoeffizienten das vorgegebene Signifikanzniveau ($p < .05$) erreichte und die Reliabilitätsschätzung mit $r_{tt} = .56$ niedrig ausfiel, wurden in der Feldstudie fast durchgehend signifikante Trennschärfen ermittelt, und die Reliabilitätsschätzung entsprach mit $r_{tt} = .84$ dem bei Persönlichkeitstests üblichen Niveau. Durch Itemselektion konnte in beiden Fällen die Reliabilität nur geringfügig gesteigert werden. Die in der Laborstudie errechnete Interraterreliabilität für den Score aggregierter Realkennzeichen betrug $r = .98$. In der Feldstudie wurde über eine logistische Regression ein Cut-Off-Wert errechnet, der von psychologischen Gutachtern als wahr eingeschätzten Aussagen von als falsch beurteilten trennte. Dieser Cut-Off-Wert lag zwischen fünf und sechs erfüllten Realkennzeichen. Unter der experimentellen Manipulation des Wahrheitsgehaltes von Aussagen in der Laborstudie konnte die Validität der merkmalsorientierten Inhaltsanalyse nur partiell bestätigt werden. Diese trennte zwar wahre und erfundene Aussagen signifikant voneinander, nicht aber wahre Aussagen von Täuschungen mit realem Erlebnishintergrund.

Schlüsselwörter: Realkennzeichen, Zeugenaussage, Glaubhaftigkeit, Reliabilität, interne Konsistenz, Validität, merkmalsorientierte Inhaltsanalyse, Kriminalpsychologie, Rechtspsychologie

Psychometric Characteristics of the Criteria Based Content Analysis

Summary

The present article combines two studies on psychometric characteristics of the Criteria Based Content Analysis (CBCA) used in the psychological assessment of testimonies. The first study was part of a larger laboratory study on the validity of content analytic judgments of the credibility of testimonies. This study was conducted with 60 subjects who were assigned to different conditions of true und false testimonies. The second study, a field study, covered 138 expert reports from a forensic psychological institution. An item analysis according to the model of the classical test theory, with reliability estimation for the score of the aggregated credibility criteria, was made with the CBCA items in both studies. The results varied considerably: Whereas only a part of the discriminative power coefficient calculated for the credibility criteria reached the given significance level ($p<.05$) and the reliability estimation was low with $r_{tt} = .56$ in the laboratory study, almost all discriminative powers in the field study were significant and the reliability estimation matched with $r_{tt} = .84$ a level which is common for personality inventories. Item selection only slightly increased the reliability in both studies. In the laboratory study, the interrater reliability for the score of aggregated credibility criteria was $r = .98$. In the field study a cut-off value was calculated by logistic regression which discriminated between testimonies assessed as true by forensic experts and testimonies assessed as false by forensic experts. The cut-off value for fulfilled credibility criteria was between five and six. In the laboratory study, the CBCA validity was only partly confirmed while the truth of the testimonies varied: True and completely fictitious testimonies were significantly different, whereas true and partly fictitious testimonies based on actual experience were not.

Keywords: Scoring the Criteria Based Content Analysis – Reliability - Validity

1 Einleitung

1.1 Rechtslage

Mit dem Urteil des Bundesgerichtshofes vom 30. Juli 1999 (BGH 30.7.1999 1 StR 618 / 98 abgedruckt in NJW 1999, S. 2746 ff.) wurde die merkmalsorientierte Inhaltsanalyse, über die die Erlebnisbegründetheit einer Aussage bestimmt werden soll, zur beweiskräftigen diagnostischen Methode im Rahmen aussagepsychologischer Begutachtung erklärt. Diese Methode, die der so genannten Undeutsch-Hypothese (Undeutsch, 1967) folgt, wonach sich wahre von falschen Aussagen anhand inhaltlicher Merkmale unterscheiden lassen, wurde durch das BGH-Urteil zum verbindlichen Bestandteil aussagepsychologischer Diagnostik. Nicht erst vor Gericht, sondern bereits im Ermittlungsverfahren kann es notwendig sein, den Glaubhaftigkeitsgehalt einer Aussage festzustellen. Auch hier sind die Grundlagen der merkmalsorientierten Inhaltsanalyse anzuwenden (Adler, 2009, 65).

1.2 Theoretischer Hintergrund

Die Aussageanalyse führt zu einer Beurteilung der Qualität einer Aussage (Volbert & Steller, 2009, 825; Volbert & Dahle, 2010). Eine Gesamtanalyse im Sinne einer Qualitätsanalyse erfordert, dass auch Teile des gesamten Begutachtungsprozesses für sich genommen eine gute Qualität aufweisen müssen. Bei der Urteilsbildung ist die Reliabilität eine Voraussetzung für die Validität, dabei ist die Qualifikation des Beurteilers eine wesentliche Komponente, was Bond und DePaulo (2008) in einer Metaanalyse mit 206 Artikeln herausgearbeitet haben.
Unter dem übergeordneten Konstrukt der Erlebnisbegründetheit sind seit Undeutschs (1967) erster Formulierung verschiedene Kataloge so genannter Realkennzeichen der Aussage vorgelegt worden, durch die häufige Verwendung in Evaluationsstudien kann aber die von Steller und Köhnken (1989) erarbeitete Fassung als die zur Zeit anerkannte bezeichnet werden. Diese Fassung ist stets gemeint, wenn im Folgenden von „Realkennzeichen" ohne weitere Bestimmung die Rede ist.
Eine diagnostische Entscheidung ist nur auf der Grundlage aggregierter Realkennzeichen zulässig, da die Erfüllung einzelner Realkennzeichen eine nur schwach gegen den Zufall gesicherte Diskriminationsschärfe bei der Kontrastierung wahrer und falscher Aussagen erkennen lässt (Fiedler & Schmid, 1999; s. auch Bundesgerichtshof, 2000). Insoweit unterscheiden sich die Anwendungsbedingungen der Realkennzeichen nicht grundsätzlich von jenen, die für psychometrische Tests gelten. Wie bei diesen ist die Aggregation von Items zur Begründung einer diagnostischen Entscheidung nur zu vertreten, wenn die aggregierten Items hinreichend homogen sind, eine Bedingung, die in der klassischen Testtheorie gewöhnlich über die Prüfung der inneren Konsistenz geklärt wird. Denn bei heterogenen Varianzquellen eines Testscores können sich die latenten Merkmalsausprägungen gegenseitig aufheben, der ermittelte Score würde so uninterpretierbar. Für die Realkennzeichen wird mit dem Begriff der Erlebnisbegründetheit ein homogen erscheinendes Konstrukt reklamiert, bei der Ausformulierung der Kennzeichen wird aber auf unterschiedliche Funktionen des Berichtsstiles Bezug genommen (Steller & Köhnken, 1989), so dass die Homogenität der Realkennzeichen trotz des eindeutig anmutenden Konstruktes der Erlebnisbegründetheit offen erscheint. Die bisherigen Bemühungen um die Evaluation der Realkennzeichen wurden vom Bestreben beherrscht, die Validität der auf die Realkennzeichen gegründeten diagnostischen Gesamtentscheidungen nachzuweisen. Die Fra-

ge nach der Stimmigkeit der Urteilsbildung, eigentlich eine Voraussetzung der Validität, ist dabei im Hintergrund geblieben.
Die innere Stimmigkeit der Urteilsbildung in der merkmalsorientierten Inhaltsanalyse von Zeugenaussagen zu beleuchten, ist Ziel der Untersuchungen, über die hier berichtet wird. In den bisher bekannt gewordenen Arbeiten zu diesem Thema wurde die Stimmigkeit der Urteilsbildung in unterschiedlicher Weise operationalisiert. Den klassischen Weg über Itemanalyse und Reliabilitätsbestimmung wählte Hommers (1997) mit einem Datensatz, der bei Kindern experimentell erhoben worden war. Er errechnete für den Score erfüllter Realkennzeichen ein Cronbachs Alpha von r_{tt} = .77. Nach Selektion eines Items mit negativer Trennschärfe – Eingeständnis von Erinnerungslücken – verbesserte sich die Reliabilitätsschätzung auf r_{tt} = .81. Die Trennschärfen der einzelnen Items waren mit Ausnahme des genannten Items signifikant positiv und erreichten Beträge von r_{it} = .20 bis r_{it} = .72. Einen anderen klassischen Weg zur Reliabilitätsbestimmung bei den Realkennzeichen, den der Re-Test-Reliabilität, wählten Horowitz at al. (1997), indem sie dieselben Aussagentranskripte zu zwei verschiedenen Zeitpunkten beurteilen ließen. Die Stabilitäskoeffizienten der drei Beurteiler reichten von r_{tt} = .85 bis r_{tt} = .91.

Rückschluss auf die innere Stimmigkeit des diagnostischen Entscheidungsprozesses gestattet auch die Treffsicherheit der resultierenden diagnostischen Gesamtentscheidungen. Dieses Postulat ergibt sich aus dem oben angesprochenen Zusammenhang zwischen Reliabilität und Validität im Modell der klassischen Testtheorie. Insoweit können auch die inzwischen zahlreich gewordenen Validitätsstudien zu den Realkennzeichen als aufschlussreich gelten (für einen Überblick siehe Niehaus, 2001; Vrij, 2005), wiewohl hier selbstredend kein Maß für die Zuverlässigkeit in einer Reliabilitätsschätzung zu erhalten ist. Belegen die nachgewiesenen Unterschiede in der Häufung der Realkennzeichen zwischen wirklich Erlebtem und erfundenen oder nacherzählten Geschichten insoweit zumindest einen homogenen Kernbestand in der Reihe der Realkennzeichen, so wurde in einem Teil der Untersuchungen bei Kontrastierung auf Itemebene deutlich, dass es einem Teil der Kennzeichen an Diskriminationsschärfe mangelt, dass damit der Bezug zum übergeordneten Konstrukt der Erlebnisbasiertheit fragwürdig bleibt (Steller at al., 1992; Krahé & Kundrotas, 1992; Lamb at al., 1997; Niehaus, 2001). Als problematisch in dieser Hinsicht müssen insbesondere jene Realkennzeichen angesehen werden, die sich auf die Aussagemotivation beziehen (Steller at al., 1992; Niehaus, 2001).

Einen weiteren Zugang zur inneren Konsistenz der diagnostischen Urteilsbildung bietet die Überprüfung des Zusammenhanges zwischen der Feststellung einzelner Realkennzeichen und der globalen Einschätzung der Erlebnisbegründetheit der Aussage durch den selben Beurteiler. Soweit das Urteil über die Erlebnisbegründetheit allein auf die Summe erfüllter Realkennzeichen gestützt wird, ist dieses Verfahren mit dem Vorgehen bei Itemanalyse und Reliabilitätsschätzung identisch. In die gutachterliche Beurteilungspraxis fließen aber weitere Informationsquellen ein, so dass der Vergleich zwischen diagnostischem Urteil und Bewertung der einzelnen Realkennzeichen mehr als eine Tautologie darstellt und über die klassische Reliabilitätsschätzung hinausgeht. Dieser Zugang wurde erstmals von Littmann und Szewczyk (1983) gewählt mit dem Ergebnis, dass den einzelnen Realkennzeichen deutlich unterschiedliches Gewicht im globalen diagnostischen Urteil über die Glaubhaftigkeit einer Aussage zukommt. Die Vorgehensweise von Littmann und Szewczyk wirft die Frage nach einem Schwellenwert in der Häufigkeit erfüllter Realkennzeichen auf, der erlebnisbegründete von nicht erlebnisbegründeten Aussagen trennt. Nach Kenntnis der Verf. hat allein Arntzen (1983, S. 22) hier eine klare Festlegung ge-

troffen – allerdings ohne Bezugnahme auf eine wissenschaftlich kontrollierte Validitätsstudie; seiner Beobachtung nach zeichnen sich wahre Aussagen durch mindestens drei Glaubhaftigkeitskriterien gemäß dem von Arntzen selbst entwickelten Kriterienkatalog aus. In späteren Veröffentlichungen (u. a. Greuel at al., 1998, Steller & Volbert, 1999) betrachtet man diesen Weg diagnostischer Entscheidungsfindung skeptisch bzw. hält ihn mit Bezug darauf, dass die Realkennzeichen unterschiedliche Funktionsweisen repräsentieren, methodisch für unzulässig. Wenn man sich auf diesen skeptischen Standpunkt begibt, sieht man sich allerdings mit der Frage konfrontiert, wie man es mit dem Aggregationsprinzip in der diagnostischen Entscheidung anhand der Realkennzeichen hält. Entschließt man sich zu einer globalen Einschätzung der Erlebnisbegründetheit einer Aussage aufgrund aggregierter Realkennzeichen, kommt man an der Frage eines wie auch immer spezifizierten Schwellenwertes nicht vorbei; selbst wenn dieser Schwellenwert alleine nicht das diagnostische Gesamtergebnis determiniert.

Nicht zuletzt hängt die interne Konsistenz diagnostischer Urteilsbildung von der Objektivität ab, mit der das zugrunde liegende Datenmaterial erfasst wird. Dieser Frage gingen verschiedene Autoren nach, indem sie die Interrater-Übereinstimmung, meist in Form von Kappa-Koeffizienten, bei der Beurteilung der Realkennzeichen errechneten. Das Übereinstimmungsmaß Kappa variierte in diesen Untersuchungen beträchtlich, so bei Anson et al. (1993) zwischen - .30 und 1, bei Horowitz et al. (1997) zwischen 0 und .75, bei Niehaus (2001) zwischen .18 und 1. Für die aggregierten Realkennzeichen errechneten Horowitz et al. (1997) als Übereinstimmungsmaß Korrelationskoeffizienten von r = .78 bis r = .89, Niehaus (2001) einen Korrelationskoeffizienten von r = .96.

1.3 Fragestellung

Die nachfolgend dargestllten Untersuchungen greifen die in den zitierten Studien bearbeiteten Probleme auf, zum Teil um erhobene Befunde zu replizieren und damit weiter empirisch abzusichern, zum Teil um sie aus veränderter Perspektive zu beleuchten. Zu replizieren sind die Befunde über die Reliabilität von Scores, die durch die Aggregation von Realkennzeichen gebildet werden, und über die Trennschärfe der einzelnen Realkennzeichen, ebenso auch die Befunde zur Validität der Entscheidung zwischen „wahr" und „unwahr" aufgrund aggregierter Realkennzeichen. Die Notwendigkeit veränderter Perspektiven ergibt sich zunächst daraus, dass die gegenwärtige Befundlage zur Brauchbarkeit der Realkennzeichen im Wesentlichen aus Laborstudien resultiert, vorwiegend aus solchen, die bei Kindern durchgeführt wurden, ferner daraus, dass die Frage nach einem Cut-Off-Wert, durch den als erlebnisbegründet und als nicht erlebnisbegründet befundene Aussagen getrennt wurden, bisher in der Forschung offenbar umgangen worden ist, wiewohl sie sich in der diagnostischen Entscheidungspraxis bei Anwendung des Aggregationsprinzipes zwangsläufig stellt. Häufig wird hierbei lediglich eine implizite Aggregation durchgeführt, indem ein Gesamturteil gefällt wird, ohne dass der Anteil der Realkennzeichen explizit gemacht wird. Es soll hier ausdrücklich die Bedeutung der Realkenneichen für das Gesamturteil geprüft werden.

2 Methode

2.1 Stichproben

2.1.1 Laborstudie

Die Daten der Laborstudie wurden im Rahmen eines größeren Projektes zur Validierung der Realkennzeichen nach Steller und Köhnken (1989) und neu formulierter mutmaßlicher Lügenkennzeichen (Hettler, 2006; Lafrenz, 2006; Maier, 2006) erhoben. Für die Item- und die Konsistenzanalyse wurden die Daten von drei Versuchs- und Kontrollgruppen mit 20 weiblichen und 40 männlichen Versuchspersonen aggregiert. Als Versuchspersonen waren Studierende der Hochschule der Polizei Villingen-Schwenningen und des Fachbereichs Betriebswirtschaft der Berufsakademie Villingen-Schwenningen angeworben worden. Das Alter der Versuchspersonen variierte zwischen 20 und 38 Jahren mit einem Median von 28 Jahren.

2.1.2 Feldstudie

Dieser testkritischen Analyse liegen die Daten von 138 Glaubhaftigkeitsbegutachtungen zugrunde, die zwischen 2000 und 2005 von insgesamt 14 Sachverständigen der Gesellschaft für wissenschaftliche Gerichts- und Rechtspsychologie GWG München durchgeführt wurden [1]. Berücksichtigt wurden nur solche Gutachten aus einer Gesamtzahl von 408 bearbeiteten Fällen, denen eine vollständige Inhaltsanalyse anhand der Realkennzeichen nach Steller und Köhnken (1989) zugrunde lag. Die Probanden hatten also jeweils einen umfassenden Spontanbericht über das fragliche Geschehen geliefert. Das Datenmaterial enthielt außerdem jeweils eine Einschätzung der Sachverständigen darüber, ob die betreffende Aussage wahrscheinlich erlebnisfundiert oder wahrscheinlich nicht erlebnisfundiert gewesen ist. In 121 der erfassten 138 Fälle handelte es sich um Zeuginnen. Das Alter der Zeuginnen und Zeugen zur Zeit der analysierten Aussage reichte von 5 bis 62 Jahren mit dem Median bei 15 Jahren. Gegenstand der Aussage war in 126 Fällen ein Tatvorwurf aus dem Bereich der Delikte gegen die sexuelle Selbstbestimmung (§§ 174 ff. StGB), in 12 Fällen ein anderes Gewaltdelikt.

2.2 Die Datenerhebung

2.2.1 Laborstudie

Die Versuchspersonen wurden aufgefordert in einem aussagepsychologischen Interview ein belastendes Ereignis im Rahmen eines Polizeieinsatzes zu berichten. Die erste Gruppe der Versuchspersonen, Studierende der Hochschule der Polizei, sollten ihrem Bericht eine eigene Erfahrung im Dienst zugrunde legen. Eine zweite Gruppe, ebenfalls Studierende der Hochschule der Polizei, sollte über ein belastendes Ereignis aus dem Polizeidienst berichten, das ihnen per Text oder Videoband übermittelt worden war; die Versuchspersonen waren gehalten, das Ereignis im Interview möglichst überzeugend als eigenes Erlebnis wiederzugeben. Die dritte Gruppe, Studierende ohne Erfahrung im Polizeidienst,

[1] Die Verf. haben der GWG, insbesondere Herrn Dr. Salzgeber und Frau Dr. Aymanns, für die Bereitstellung der Daten zu danken.

sollte eine ausgedachte Geschichte zu einem belastenden Ereignis im Polizeidienst möglichst überzeugend als eigenes Erlebnis erzählen. Die drei Gruppen waren mit je 20 Personen besetzt, die hinsichtlich Alter und intellektuellem Niveau vergleichbar waren. Das Untersuchungsdesign folgte dem Modell, das Niehaus (2001) für ihre breit angelegte Studie zur Validität der merkmalsorientierten Inhaltsanalyse entwickelt hatte. Die Interviews wurden unter weitgehender Standardisierung des Interviewerverhaltens von Psychologiestudierenden ausgeführt, die zuvor ein entsprechendes Training absolviert hatten. Die Interviews wurden wörtlich transkribiert und von anderen trainierten Studierenden der Psychologie, die die Entstehungsbedingungen der einzelnen Interviews nicht kannten, hinsichtlich der Erfüllung von Realkennzeichen im Sinne von Steller und Köhnken (1989) ausgewertet. Die Versuchsbedingungen der Laborstudie führten dazu, dass zwei Realkennzeichen gegenstandslos wurden: Das vorrangig auf kindliche Zeugen bezogene Kriterium „phänomengemäße Schilderung unverstandener Handlungselemente" besitzt bei der Population der Studie keine Relevanz; dem Kriterium „deliktspezifische Aussageelemente" fehlt ein bedeutsamer Bezug zum Thema der analysierten Erlebnisschilderungen.

2.2.2 Feldstudie

Die Daten der Feldstudie wurden routinemäßig erstellten aussagepsychologischen Gutachten entnommen. Die 14 Sachverständigen hatten die von ihnen erhobenen Daten ebenfalls anhand der Realkennzeichen von Steller und Köhnken (1989) ausgewertet, die Vorgaben des Bundesgerichtshofes mit seinem Urteil vom 30.07.1999 wurden von ihnen als verbindlich erachtet. In einer Reihe von Fällen, namentlich in solchen fortgesetzten sexuellen Missbrauchs, werteten die Sachverständigen verschiedene Aussageteile separat aus. Hier wurde jeweils jene Analyse für die Datenverarbeitung herangezogen, der die Gutachter die beste Fundierung zugesprochen hatten. Ansonsten wurden die diagnostischen Urteile unmodifiziert übernommen.

2.3 Die Datenverarbeitung

Mit den in beiden Untersuchungen erhobenen Datensätzen wurden jeweils Itemanalysen nach dem klassischen Modell mit Reliabilitätsschätzungen nach Cronbach`s Alpha gerechnet. Mit den Daten der Laborstudie, die von zwei unabhängigen Ratern codiert waren, war es zudem möglich, die Interraterreliabilität zu bestimmen. Über Itemselektionen nach dem Kriterium der Trennschärfe wurden Optimierungen der Reliabilitätsschätzungen angestrebt. Die Berechnungen erfolgten auf der Grundlage einer dichotomen Codierung der Items, d. h., die einzelnen Realkennzeichen gingen mit dem diagnostischen Urteil vorhanden versus nicht vorhanden in die Datenverarbeitung ein. Angesichts der besonderen Validitätsproblematik, die die motivationsbezogenen Realkennzeichen zu belasten scheint (Steller at al., 1992; Niehaus, 2001), wurde in der Laborstudie die Itemanalyse ohne diese Merkmalsgruppe wiederholt. Mit Erfassung der globalen gutachterlichen Einschätzung der Glaubhaftigkeit im Anschluss an die kriterienbasierte Inhaltsanalyse der Aussage gestatteten es die Daten der Feldstudie, einen weiteren Aspekt der inneren Konsistenz diagnostischer Urteilsbildung bei Anwendung der Realkennzeichen zu bearbeiten. Über die Bestimmung einer logistischen Regression vom Summenscore der Realkennzeichen auf die Kategorisierung „glaubhaft" versus „ nicht glaubhaft" wurde ein Maß für

die innere Konsistenz der diagnostischen Urteilsbildung berechnet. Das Modell der logistischen Regression erlaubt zudem, die Bestimmung eines Schwellenwertes für die Klassifikation nach den Kriterien „glaubhaft" versus „nicht glaubhaft".

Mit der experimentellen Definition der unabhängigen Variablen in der Laborstudie ermöglichen die vorliegenden Daten eine weitere Validitätskontrolle der merkmalsorientierte Inhaltsanalyse. Der Wert dieser Kontrolle bemisst sich daran, dass sich die unwahre Zeugenaussage nach den Bedingungen von Erfinden einer Aussage versus Nacherzählen eines kontextfremden real erlebten Ereignisses unterscheidet. Die Bedingung der wahren Zeugenaussage kann so mit zwei Bedingungen unwahrer Aussagen kontrastiert werden. Nach den Befunden von Niehaus (2001) war hier ein Gefälle aggregierter Realkennzeichen von der wahr aussagenden Gruppe über die nacherzählende Gruppe zur Aussage erfindenden Gruppe zu erwarten.

3 Ergebnisse

3.1 Itemanalyse und Reliabilitätsschätzung – Laborstudie

Wie unter 2.2.1. ausgeführt wurde, musste die Itemanalyse mit den Daten der Laborstudie auf 17 Items beschränkt werden (Tabelle 1). Nur neun der 17 Items wiesen signifikante Trennschärfekoeffizienten (unter $p < .05$) auf. Entsprechend niedrig fiel die Reliabilitätsschätzung nach Cronbachs Alpha mit $r_{tt} = .56$ aus. Bei einer schrittweisen Selektion der Items mit den jeweils niedrigsten Trennschärfen konnte die Reliabilitätsschätzung mit dem siebten Selektionsschritt auf $r_{tt} = .61$ gesteigert werden. Die weitergehende Selektion führte wieder zur Abnahme der Reliabilität. Eine getrennte Berechnung für die drei Gruppen scheiterte an der kleinen Stichprobengröße mit jeweils 20 Probanden.
Die Itemanalyse wurde unter Weglassen problematisch erscheinender motivationsbezogener Realkennzeichen wiederholt. Entfernt wurden jene Items aus der Liste von Steller und Köhnken (1989), deren Koeffizienten in der ersten Analyse nicht das vorgegebene Signifikanzniveau von fünf Prozent erreicht hatten. Die Reliabilitätschätzung betrug nun $r_{tt} = .61$. Die Rangordnung der Trennschärfen veränderte sich dabei nicht. Die Interrater-Übereinstimmung für den Score der aggregierten Realkennzeichen wurde mit $r_{tt} = .98$ bestimmt und kann als zufrieden stellend beurteilt werden.

Tabelle 1: Itemanalyse der Realkennzeichen mit den Daten der Laborstudie (N = 60)

Realkennzeichen	**S**	**r_{it}**
Logische Konsistenz	.88	.224*
Unstrukturierte Darstellung	.33	.198
Detailreichtum	.80	.423**
Raum-zeitliche Verknüpfung	.55	.155
Interaktionsschilderungen	.43	.122
Gesprächswiedergaben	.82	.341**
Handlungskomplikationen	.50	.244*
Ausgefallene Details	.58	.273*
Nebensächliche Details	.45	.328**
Indirekt Handlungsbezogenes	.23	.253*
Eigenpsychisches	.92	-.067
Fremdpsychisches	.57	.185
Spontane Verbesserung	.33	.241*
Eingestehen von Erinnerungslücken	.43	.070
Einwände gegen eigene Aussage	.60	.222*
Selbstbelastung	.38	.088
Täterentlastung	.10	.065

S: Schwierigkeitsindex
r_{it}: Trennschärfeindex (Korrigierte Item-Total-Korrelation)

*: $p \leq .05$
**: $p \leq .01$

3.2 Itemananlyse und Reliabilitässchätzung Feldstudie

Im Unterschied zu den Ergebnissen der Itemanalyse auf der Basis der Labordaten erwiesen sich fast alle Trennschärfen der Realkennzeichen in der Feldstudie als signifikant (unter $p < .05$). Lediglich das selten besetzte Item „indirekt handlungsbezogene Elemente" erreichte nicht das vorgegebene Signifikanzniveau (Tabelle 2). Die Reliabilitätsschätzung nach Cronbachs Alpha betrug auf der Grundlage aller 19 Realkennzeichen $r_{tt} = .84$. Die Selektion von Items mit niedrigen Trennschärfekoeffizienten erbrachte nur geringfügige Steigerungen des Reliabilitätsmaßes. Beim fünften Selektionsschritt wurde eine Reliabilitätsschätzung von $r_{tt} = .854$ erreicht.

Tabelle 2: Itemanalyse der Realkennzeichen in der Feldstudie (N = 138)

Realkennzeichen	S	r_{it}
Logische Konsistenz	.64	.705**
Unstrukturierte Darstellung	.30	.471**
Detailreichtum	.67	.677**
Raum-zeitliche Verknüpfung	.59	.550**
Interaktionsschilderungen	.55	.528**
Gesprächswiedergaben	.41	.517**
Handlungskomplikationen	.41	.482**
Ausgefallene Details	.38	.315**
Nebensächliche Details	.28	.283**
Unverstandene Handlungen	.15	.293**
Indirekt Handlungsbezogenes	.01	.063
Eigenpsychisches	.60	.593**
Fremdpsychisches	.26	.548**
Spontane Verbesserung	.09	.291**
Eingestehen von Erinnerungslücken	.14	.212**
Einwände gegen eigene Aussage	.04	.188*
Selbstbelastung	.40	.408**
Täterentlastung	.44	.350**
Deliktspezifisches	.41	.538**

S: Schwierigkeitsindex
r_{it}: Trennschärfeindex (Korrigierte Item-Total-Korrelation)
*: $p \leq .05$
**: $p \leq .01$

3.3 Konsistenz zwischen Summenscore und der Schwellenwert - Feldstudie

Die Sachverständigen der Feldstudie beurteilten 79 Aussagen als glaubhaft und 59 als nicht glaubhaft. Das in der logistischen Regression von den Summenscores der Realkennzeichen auf die Zuweisung zu den beiden Kategorien errechnete R^2 nach Nagelkerke betrug .94 (Tabelle 3). Der sich hierin ausdrückende Zusammenhang zwischen Häufigkeit festgestellter Realkennzeichen und Glaubhaftigkeitsbeurteilung findet seinen Niederschlag auch in der nahezu perfekten Übereinstimmung zwischen den tatsächlichen Klassifikationen der Sachverständigen und den durch das logistische Regressionsmodell vorhergesagten Klassifikationen (Tab. 3). Im Umkehrschluss bedeutet dieses Resultat, dass die Sachverständigen bei ihrer Glaubhaftigkeitsbeurteilung einzelnen Realkennzeichen kein besonderes Gewicht beimessen. Die im Regressionsmodell zu errechnende kritische Grenze für die Glaubhaftigkeitsbeurteilung liegt zwischen fünf und sechs erfüllten Realkennzeichen. Während die Wahrscheinlichkeit für die Einschätzung „nicht glaubhaft" für den Summenscore von fünf noch p = .86 beträgt, sinkt sie für den Summenscore sechs auf p = .47. Lediglich bei drei der als nicht glaubhaft eingestuften Aussagen wurden mehr als fünf Realkennzeichen erkannt (zweimal sieben, einmal sechs) und nur bei einer

der 79 als glaubhaft eingeschätzten Aussagen wurden weniger als sechs Realkennzeichen festgestellt.

Tabelle 3: Diagnostische Klassifikation "glaubhaft" vs. "nicht glaubhaft" und ihre Verteilung um den Cut-Off-Wert des Summenscores der Realkennzeichen im Modell der logistischen Regression

Klassifikation	**Summe ≥ 6**	**≤ 5**	**N**	$\bar{x}$	**S**
Glaubhaft	78	1	79	10.0	2.4
Nicht glaubhaft	3	56	59	2.5	1.8
N	81	57	138		

Variablen der logistischen Regression:

Summenscore: B = 1.932	SE = 0.479	Chi^2 = 16.257***
Konstante: B = -11.483	SE = 2.969	Chi^2 = 14.960***
Modellanpassung nach Nagelkerke:	R^2 = 0.941	

B: Regressionskoeffizient; SE: Standardfehler von B; ***: p < .001 im Chi-Quadrat-Test nach Wald bei 1 Freiheitsgrad

3.4 Validitätskontrolle der aggregierten Realkennzeichen in der Laborstudie

Die Manipulation der Interviewbedingungen in der Laborstudie gestattete eine Kontrastierung der Häufigkeit erfüllter Realkennzeichen unter der Bedingung wahrer Aussagen mit der Aggregation von Realkennzeichen bei nacherzählten Fremderlebnissen und bei erfundenen Erlebnissen. In Anlehnung an die Ergebnisse von Niehaus (2001) war die Hypothese zu testen, dass ein Gefälle in der Aggregation von Realkennzeichen zu beobachten sei von der Bedingung wahrer Aussagen über die Nacherzählung zur erfundenen Geschichte. Prüfstatistisch lagen hier die Voraussetzungen für den Trendtest von Jonckheere vor (vgl. Bortz & Lienert, 1998, 149 ff.). Der Test führte zu einer auf dem Fünf-Prozent-Niveau signifikanten Prüfgröße (Tabelle 4). Die Gruppenkontraste mit dem U-Test von Mann-Whitney führten nur bei der Gegenüberstellung der wahren und der erfundenen Aussagen zu einem signifikanten Resultat (unter p < .05).

Tabelle 4: Summenscores der Realkennzeichen unter den Aussagebedingungen "Wahrheit",irreführende Nacherzählung" und "Erfindung" in der Laborstudie

Aussagebedingungen	N	$\bar{x}$	Md	S	Min	Max
Wahrheit	20	9.90	10.0	2.59	5	15
Nacherzählung	20	8.95	9.5	2.50	4	14
Erfindung	20	7.90	7.0	2.80	4	13

u = -2.304*

$\bar{x}$ = Mittelwert; Md = Median; S = Standardabweichung; Min = niedrigster Wert; Max = höchster Wert; u = Prüfgröße im asymptotischen Jonckheere-Test;

Gruppenkontraste (Mann-Whitney-U-Test):
Wahrheits vs. Nacherzählung - z = -1.229
Wahrheit vs. Erfindung - z = -2.205*
Nacherzählung vs. Erfindung - z = -1.161
*: $p \leq .05$

4 Diskussion

Die unter der Frage nach der psychometrischen Qualität der merkmalsorientierten Inhaltsanalyse durchgeführte Studie führt in beiden Untersuchungen zu unterschiedlichen Resultaten. Während die Feldstudie zu einer Reliabilitässchätzung für den Score erfüllter Realkennzeichen mit r_{tt} = .84 einen Betrag ergab, der im Vergleich mit den Reliabilitäten gebräuchlicher Persönlichkeitstests bestehen kann (vgl. Westhoff, 1993), erwies sich die Reliabilität in der Laborstudie mit r_{tt} = .56 als niedrig. Und sie steigerte sich nur geringfügig nach Aussonderung besonders trennschwacher Items. Beim Versuch einer Erklärung für die Divergenz beider Ergebnisse wird man zunächst an unterschiedliche Varianzverhältnisse in den Stichproben zu denken haben, die den Untersuchungen zugrunde lagen, da die Höhe von Korrelationsmaßen von der Merkmalsvarianz innerhalb einer Stichprobe abhängt. Betrachtet man jedoch die Streubreite der Scores in der Laborstudie, so kann diese Erklärung kaum befriedigen. Man kommt wohl nicht an dem Postulat vorbei, dass in der Labor- und in der Feldstudie unterschiedliche Varianzquellen die Ergebnisse bestimmt haben. Dieses Postulat berührt die grundsätzliche Frage nach der Übertragbarkeit von Ergebnissen, die im Labor gewonnen wurden, auf die gutachterliche Entscheidungspraxis. Es handelt sich um verschiedene Delikte in beiden Studien.

Das relativ hohe Reliabilitätsmaß, das für die aggregierten Realkennzeichen in der Feldstudie errechnet wurde, veranlasst zunächst zu einer optimistischen Einschätzung der psychometrischen Verwendbarkeit der merkmalsorientierten Inhaltsanalyse, zumal auf den ähnlich erfolgreich verlaufenen Versuch von Hommers (1997) verwiesen werden kann. Die günstigen Reliabilitätsmaße sagen allerdings nur etwas über die interne Qualität des Beurteilungsinstrumentes aus, nicht über jene des beurteilten Gegenstandes. Konkret bedeutet dies, dass die Anwender im vorliegenden Falle offenbar über ein klares Konzept von der Erlebnisnähe einer Aussage und ihrer Operationalisierung in den Real-

kennzeichen verfügten. Auf die Validitätsfrage gibt diese Feststellung keine Antwort. So liefert auch der relativ verlässlich erscheinende Cut-Off-Wert von fünf bzw. sechs aggregierten Realkennzeichen, der für die Trennung von wahren und unwahren Aussagen in der Feldstudie ermittelt wurde, keine valide Entscheidungsgrundlage für die gutachterliche Praxis. Er bestätigt lediglich die innere Stimmigkeit des diagnostischen Urteils.

Da Reliabilität bekanntlich eine wesentliche Voraussetzung für die Validität eines diagnostischen Verfahrens ist, sind die getroffenen Feststellungen zur inneren Stimmigkeit einer psychometrischen Verwendung der merkmalsorientierten Inhaltsanalyse jedoch keineswegs belanglos. Nachdem die Feldstudie die Validitätsfrage ausgeklammert hat, mag man sich in dieser Hinsicht mit den inzwischen zahlreich gewordenen empirischen Nachweisen begnügen, dass die Wahrscheinlichkeit für das Auftreten von Realkennzeichen bei wahren Aussagen höher ist als bei erfundenen Aussagen (Überblicke bei Steck, 2009; Greuel et al., 1998; Niehaus, 2001, Steller & Volbert, 1999; Vrij, 2005) konkrete Quellen einfügen). Man kann auch das Ergebnis unserer Laborstudie in diese Reihe einfügen. Der Optimismus in der Einschätzung der Validität der merkmalsorientierten Inhaltsanalyse hat allerdings mit der Untersuchung von Niehaus (2001) einen Dämpfer erfahren, als in den Bedingungen unwahrer Aussagen zwischen erfundenen und nacherzählten Aussagen differenziert wurde. Der Bundesgerichtshof hatte bereits in seinem Urteil vom 30.07.1999 (BGH 30.7.1999 1 StR 618 / 98 abgedruckt in NJW 1999, S. 2746 ff.) auf die entsprechende Lücke im Validitätsnachweis aufmerksam gemacht. Die Problematik ist in der vorliegenden Laborstudie wieder deutlich sichtbar geworden, da es nicht gelang, in der Aggregation erfüllter Realkennzeichen einen überzufälligen Unterschied zwischen Eigenerlebnissen und nacherzählten Erlebnissen nachzuweisen. Aber auch der Unterschied zwischen wirklichen und erfundenen Erlebnissen fiel trotz der signifikanten Testgröße im Hinblick auf die Einzelfalldiagnostik nicht sehr überzeugend aus; schließlich überschnitten sich die Häufigkeiten erfüllter Realkennzeichen zwischen beiden Gruppen beträchtlich. Diese Feststellung unterstreicht die Problematik der Übertragung von Laborbefunden auf die diagnostische Entscheidungspraxis. Darüber hinaus sehen wir wie Niehaus (2008) ein weiteres Defizit bei der Unterscheidung zwischen erlebten und erfunden Aussagen darin, dass verbale Lügenmerkmale bis heute nicht bekannt sind.

Literatur

Adler, F. (2009). Rechtsfragen bei Glaubhaftigkeitseinschätzungen – Aussageanalyse und Polygraphentest. In M. Hermanutz & S. Litzcke (Hrsg.), Vernehmung in Theorie und Praxis (55 – 68). Stuttgart: Boorberg.

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

Arntzen, F. (1983). Psychologische Beurteilung der Glaubwürdigkeit von Zeugenaussagen. In F. Lösel (Hrsg.), *Kriminalpsychologie* (173 – 180). Weinheim: Beltz - Psychologie Verlags Union.

Bond, C. F. Jr. & DePaulo, B. M. (2008). Accuracy of Deception Judgments. *Pers Soc Psychol Rev* 2006 10: 214-234

Bortz, J. & Lienert, G. A. (1998). *Kurzgefasste Statistik für die klinische Forschung.* Berlin: Springer.

Bundesgerichtshof (2000). Anforderungen an Glaubhaftigkeitsgutachten. *Neue Zeitschrift für Strafrecht,* 2000 (2), 100 – 105.

Fiedler, K. & Schmid, J. (1999). Gutachten über Methodik und Bewertungskriterien für psychologische Glaubwürdigkeitsgutachten. *Praxis der Rechtspsychologie,* 9, 5 – 45.

Greuel, L., Offe, S., Fabian, A., Wetzels, P., Offe, H. & Stadler, M. (1998). *Glaubhaftigkeit der Zeugenaussage.* Weinheim: Psychologie Verlags Union.

Hettler, S. (2006). *Wahre und falsche Zeugenaussagen.* Saarbrücken: VDM Verlag Dr. Müller.

Hommers, W. (1997). Die aussagepsychologische Kriteriologie unter kovarianzstatistischer und psychometrischer Perspektive. In L. Greuel, T. Fabian & M. Stadler (Hrsg.), *Psychologie der Zeugenaussage* (87 – 100). Weinheim: Psychologie Verlags Union.

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

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

Lafrenz, B. (2006). *Wahrheit und Lüge bei Zeugenaussagen.* Saarbrücken: VDM Verlag Dr. Müller.

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

Littmann, E. & Szewczyk, H. (1983). Zu einigen Kriterien und Ergebnissen forensisch-psychologischer Glaubwürdigkeitsbegutachtungen von sexuell missbrauchten Kindern und Jugendlichen. *Forensia,* 4, 55 – 72.

Maier, B. (2006). *Glaubhaftigkeitsdiagnostik von Zeugenaussagen.* Saarbrücken: VDM Verlag Dr. Müller.

Niehaus, S. (2001). *Zur Anwendbarkeit inhaltlicher Glaubhaftigkeitsmerkmale bei Zeugenaussagen unterschiedlichen Wahrheitsgehaltes.* Frankfurt/Main: Peter Lang.

Niehaus, S. (2008). Merkmalsorientierte Inhaltsanalyse. In: Renate Volbert und Max Steller (Hrsg.). *Handbuch der Rechtspsychologie* (311 – 321). Göttingen: Hogrefe.

Steck, P. (2009). Aussageanalyse. In M. Hermanutz & S. Litzcke (Hrsg.), Vernehmung in Theorie und Praxis (55 – 68). Stuttgart: Boorberg.

Steller, M. & Köhnken, G. (1989). Criteria-based statement analysis. In D. C. Raskin (Ed.), *Psychological methods for investigation and evidence* (217 – 245). New Y-ork: Springer.

Steller, M. & Volbert, R. (1999). Wissenschaftliches Gutachten. Forensisch-aussagepsychologische Begutachtung (Glaubwürdigkeitsbegutachtung). *Praxis der Rechtspsychologie*, 9, 46 – 112.

Steller, M., Wellershaus, P. & Wolf, T. (1992). Realkennzeichen in Kinderaussagen: Empirische Grundlagen der Kriterienorientierten Aussageanalyse. *Zeitschrift für experimentelle und angewandte Psychologie*, 39, 151 – 170.

Undeutsch, U. (1967). Beurteilung der Glaubhaftigkeit von Aussagen. In U. Undeutsch (Hrsg.), Handbuch der Psychologie, Band 11, *Forensische Psychologie* (26 – 181). Göttingen: Hogrefe.

Volbert, R. & Steller, M. (2009). Die Begutachtung der Glaubhaftigkeit. In K. Foerster & H. Dressing (Ed.). *Psychiatrische Begutachtung* (693 – 728). München: Elsevier.

Volbert, R & Dahle, K.-P. (2010). *Forensisch – psychologische Diagnostik im Strafverfahren*. Göttingen: Hogrefe,

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

Westhoff, G. (1993). *Handbuch psychosozialer Messinstrumente.* Göttingen: Hogrefe.