Aus dem Reformstudiengang Medizin
der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

**DISSERTATION**

# Validierung des „Berliner Global Rating“ (BGR) - ein Instrument zur Prüfung kommunikativer Kompetenzen Medizinstudierender im Rahmen klinisch-praktischer Prüfungen (OSCE)

zur Erlangung des akademischen Grades
Doctor rerum medicarum (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von
Dipl.-Psych. Simone Scheffer
aus Berlin

Gutachter: 1. Prof. Dr. med. W. Burger
2. Prof. Dr. B. Babitsch
3. Prof. Dr. phil. D. Kleiber

Datum der Promotion: 29.06.2009

# Inhaltsverzeichnis

# 1 Einleitung

## 1.1 Ärztliche Gesprächsführung - Relevanz des Themas

Ein Arzt[1] erhebt in seinem Berufsleben ungefähr 200.000 Anamnesen (1). Dazu kommen zahlreiche weitere Formen des ärztlichen Gesprächs, das zentraler Bestandteil des diagnostischen und therapeutischen Prozesses ist. Neben dem Wissen um Behandlungsmethoden und Krankheitsbilder gehört das Gespräch mit Patienten damit zu den elementaren, nicht delegierbaren ärztlichen Fachkompetenzen (2).

Die Qualität der ärztlichen Kommunikation steht seit längerem in der Kritik. Häufige Mängel und Fehler in der Arzt-Patienten-Kommunikation, denen unzureichende kommunikative Kompetenzen des Arztes zugrunde liegen, sind in verschiedenen Übersichtsarbeiten im Detail beschrieben (1-4). In einer vergleichenden Untersuchung, in der in acht europäischen Ländern die kommunikativen Kompetenzen von Ärzten und anderen Professionellen im Gesundheitswesen durch Patienten beurteilt wurden, liegt Deutschland auf dem vorletzten Platz (5). In deutschen hausärztlichen Praxen sind die Gespräche im europäischen Vergleich am kürzesten, am stärksten medizinisch und am wenigsten an den Bedürfnissen des Patienten orientiert (6). Wird das ärztliche Gespräch als wichtigstes diagnostisches Instrument nicht optimal genutzt, werden häufig unnötige und kostenintensive apparative Untersuchungen angeordnet. Unzufriedenheit mit dem Arzt aufgrund unzureichender Kommunikation spielt darüber hinaus eine große Rolle in der Entscheidung von Patienten, bei ungünstigem Therapieverlauf eine Klage einzureichen. In einer prospektiven Studie in den USA zeigte sich, dass die Patienten von Ärzten, die bei der Prüfung ihrer kommunikativen Fähigkeiten besonders schlecht abgeschnitten hatten, sich später signifikant häufiger bei medizinischen Aufsichtsorganen („medical regulatory authorities“) über ihre Ärzte beschwerten (7).

Für die vielfältigen positiven Auswirkungen gelungener Arzt-Patienten-Kommunikation gibt es zahlreiche empirische Belege, deren vollständige Darstellung den Rahmen dieser Arbeit sprengen würde. Effekte manifestieren sich bei Patienten (und Ärzten) auf psychologischer, verhaltensbezogener und somatischer Ebene sowie in der Qualität der Gesundheitsversorgung: Verschiedene Aspekte des ärztlichen Kommunikationsstils, die Bestandteile eines „Patientenzentrierten Gesprächsstils“ sind, gehen mit einer höheren Zufriedenheit mit der Behandlung (8, 9) sowie Angst- und Stressreduktion auf Patientenseite (10) einher. Im Hinblick auf die Behandlungsverläufe und Heilungsprozesse wurden verschiedene positive Folgen beschrieben.

[1] Es sind immer beide Geschlechter gemeint, aus Gründen der Lesbarkeit wird darauf verzichtet, beide zu nennen.

So ist eine patientenzentrierte Gesprächsführung u.a. mit einer korrekteren Identifikation der Probleme des Patienten durch den Arzt (11), mit der Verschreibung von weniger Medikamenten (9) sowie weniger Überweisungen assoziiert (12). Unter patientenzentrierter Gesprächsführung wird ein kooperativer Interaktionsstil verstanden, der es dem Patienten ermöglicht, sich aktiv am Gespräch, der Entscheidungsfindung und der Therapieplanung zu beteiligen. Wichtige Elemente sind die Erfassung des subjektiven Krankheitsverständnisses des Patienten, die Förderung der Arzt-Patienten-Beziehung im Sinne einer therapeutischen Allianz und eine Grundhaltung, die die Wahrnehmung des Patienten als Person erlaubt (12, 13).

In verschiedenen Übersichtsarbeiten sind Studien, die Effekte gelungener Arzt-Patienten-Kommunikation untersuchen, zusammenfassend bewertet (1, 14, 15). Eine patientenzentrierte Gesprächsführung ist routinemäßig im klinischen Alltag durchführbar, ohne Gespräche unbotmäßig zu verlängern, vorausgesetzt, dass der Arzt die relevanten Techniken beherrscht (9, 10).

### 1.2 Konsequenzen für die medizinische Ausbildung

Bis in die 80er Jahre erfuhr das Thema „ärztliche Gesprächsführung“ in der medizinischen Ausbildung wenig Aufmerksamkeit. Dem lagen im Wesentlichen vier verschiedene Annahmen zugrunde: Eine Annahme bestand darin, dass ärztliche Gesprächsführung nicht gelehrt werden müsste, da Medizinstudierende kommunikative Kompetenzen bereits besäßen. Eine zweite Annahme besagte, dass diese Fertigkeiten nicht explizit gelehrt werden müssten, da sie während des Studiums en passant gelernt würden. Eine dritte Gruppe von Kritikern bezweifelte generell den Nutzen von kommunikativen und sozialen Kompetenzen, die ihnen im Vergleich zu medizinischem Fachwissen unwichtig schienen (16). Ein viertes Argument besagte, dass kommunikative Kompetenzen nicht erlernbar seien (17).

Durch die Forschungsergebnisse zu den Mängeln in der Arzt-Patienten-Kommunikation und den positiven Effekten gelungener Kommunikation fand das Thema Kommunikation in der Medizin in den neunziger Jahren Eingang in die Ausbildungsforschung und die Curriculumsplanung (16). Studien untersuchten, ob und wie ärztliche Gesprächsführung in der medizinischen Ausbildung gelehrt werden kann. Die Ergebnisse zeigten, dass kommunikative und soziale Kompetenzen erlernbare Fertigkeiten sind, die durch Training verbessert werden können; sie werden jedoch nicht regelhaft als „Nebenprodukt“ des Studiums erworben (18, 19). Die Ausbildung in patientenorientierter Kommunikation sollte unter Berücksichtigung ihrer Bedeutung für die ärztliche Tätigkeit gründlich und umfassend sein (1). Als besonders wirksam hat sich ein

regelmäßiger, hochstrukturierter, interaktiver Unterricht in Kleingruppen mit einem hohen Übungsanteil und qualitativ hochwertigem Feedback erwiesen. Feedback ist ein essentielles Element erfolgreicher Kommunikationstrainings. Feedback bezeichnet dabei eine konstruktive Rückmeldung an den Lernenden in strukturierter Form und nach vorgegebenen Regeln. In mehreren Übersichtsarbeiten ist die Evidenz zur Effektivität von Kommunikationstrainings zusammengefaßt dargestellt (2, 17, 20, 21).

Für die Vermittlung kommunikativer Kompetenzen ist die Arbeit mit Simulationspatienten besonders fruchtbar: Simulationspatienten (SPs) sind speziell ausgebildete Laien und/oder Schauspieler, die Patientenrollen realistisch und standardisiert darstellen können. Zur Simulation gehören neben der Präsentation der Erkrankungssymptome auch die Wiedergabe der Patientengeschichte sowie der Persönlichkeitsmerkmale der Rolle. Die Methode „Simulationspatient“ wurde in den 60er Jahren entwickelt (22). Gegenüber dem Einsatz von realen Patienten verbindet sie verschiedene Vorteile: Studierende haben die Möglichkeit, ihre kommunikativen Kompetenzen in einem geschützten Rahmen zu trainieren. Reale Patienten werden nicht belastet; unzumutbare Situationen wie beispielsweise die Überbringung schwerwiegender Diagnosen können mit SPs geübt und sogar wiederholt trainiert werden. Dabei kann der Schwierigkeitsgrad der Gespräche dem Ausbildungsstand angepaßt werden. SP sind außerdem darin ausgebildet, strukturiert Feedback zu dem Gespräch zu geben - ein für das Lernen entscheidender Vorteil gegenüber realen Patienten (23).

Im Englischen wird oftmals zwischen „Communications Skills“ und „Interpersonal Skills“ unterschieden. Erstere schließen die Anamneseerhebung, die Erläuterung von Diagnose und Prognose sowie Therapieempfehlungen und Aufklärung ein. Letztere beziehen sich auf die Beziehung zwischen Arzt und Patient und den Prozess der Gesprächsführung, sie beinhalten respektvolles Verhalten, Aufmerksamkeit und Wertschätzung (24). In der deutschsprachigen medizinischen Ausbildungsforschung ist diese Differenzierung nicht üblich, hier wird von kommunikativen Kompetenzen gesprochen.

### 1.2.1 Internationale Ausbildungssituation

Im angloamerikanischen Raum wurden beginnend in den 90er Jahren in verschiedenen Ländern Empfehlungen und Richtlinien für die medizinische Ausbildung entwickelt. Darin werden Kompetenzen beschrieben, die Absolventen aufweisen sollen. Kompetenzen beinhalten ein fächerübergreifendes Repertoire an Wissen, Fertigkeiten und Haltungen und beziehen sich auf

das „Endprodukt“ der medizinischen Ausbildung. In allen Richtlinien werden kommunikative und soziale Kompetenzen für die Ausbildung gefordert, meist wird auch deren Überprüfung empfohlen.

Das General Medical Council in Großbritannien beispielsweise spricht sich in seinen Empfehlungen dafür aus, Kommunikationstrainings als ein Pflichtfach vom ersten Studienjahr an einzuführen, welches sich durch das gesamte Studium ziehen und geprüft werden sollte. „Tomorrow’s Doctors“ beschreibt, welche kommunikativen Kompetenzen Absolventen aufweisen müssen und fordert, dass den Studierenden Gelegenheit gegeben werden muss, diese zu erwerben (25, 26). Die britischen Universitäten haben daraufhin ihre Curricula überarbeitet. In einer postalischen Befragung aller 26 medizinischen Fakultäten in Großbritannien zeigte sich 1998 jedoch noch eine große Heterogenität in der curricularen Umsetzung der Empfehlungen. Nur die Hälfte der Fakultäten in Großbritannien hatte bis zu diesem Zeitpunkt verpflichtende Prüfungen kommunikativer Kompetenzen eingeführt (16).

Das Liaison Committee on Medical Education (LCME) und das Committee on Accreditation of Canadian Medical Schools (CACMS), die Akkreditierungsbehörden für die ärztliche Ausbildung in den USA bzw. Kanada, schreiben als Voraussetzung für die Akkreditierung Unterricht in Gesprächsführung in verschiedenen ärztlichen Kontexten vor (27). Das Accreditation Council for Graduate Medical Education (ACGME) definierte sechs übergeordnete, miteinander verknüpfte Kompetenzbereiche für die ärztliche Weiterbildung, von denen einer kommunikative Kompetenz ist. Für die Prüfung dieser Kompetenzen wurden ebenfalls Empfehlungen entwickelt (28).

In Kanada wurde 1993 das „CanMEDS 2000 Projekt“ initiiert mit dem Ziel, ein den Herausforderungen des Gesundheitswesens entsprechendes Kompetenzprofil für die ärztliche Weiterbildung zu entwickeln. Das „framework for competency-based education“ definiert sieben verschiedene sogenannte „Rollen“ des Arztes; diese beinhalten Schlüsselkompetenzen und Ausbildungsziele, die für alle Fachrichtungen maßgeblich sind. Zu den Anforderungen gehört die Rolle des „Communicators“, die dezidiert notwendige kommunikative Kompetenzen beschreibt (29, 30).

Zwei weitere Richtlinien beschäftigen sich ausschließlich mit dem Bereich kommunikative Kompetenzen und geben Empfehlungen für die Ausbildung: das Toronto Consensus Statement (2) und das Kalamazoo Consensus Statement (31).

In Kontinentaleuropa sind niederländische und schweizerische Universitäten Vorreiter in der intensiven Aus- und Weiterbildung im Bereich kommunikativer und sozialer Kompetenzen, z. B. die Universitäten Maastricht und Basel (18, 32).

Zusammenfassend ist festzuhalten, dass alle o. g. Empfehlungen und Richtlinien für die medizinische Aus- bzw. Weiterbildung zu übereinstimmenden Ergebnissen kommen: Kommunikative Kompetenzen sind in den Fokus der medizinischen Ausbildung gerückt und werden als Kernkompetenz definiert. Strukturierte Unterrichtsangebote werden in allen genannten Publikationen explizit gefordert und in der Folge an vielen Fakultäten umgesetzt. Daraus ergibt sich auch die Forderung nach der Überprüfung des Kompetenzerwerbs, die vielerorts schon Standard ist.

### 1.2.2 Ausbildungssituation in Deutschland

Im internationalen Vergleich haben die medizinischen Fakultäten in Deutschland großen Nachholbedarf im Hinblick auf die Vermittlung kommunikativer Kompetenzen in der medizinischen Ausbildung. In Absolventenstudien werden übereinstimmend kommunikative und soziale Kompetenzen als elementare Voraussetzungen für die ärztliche Berufsausübung benannt und gleichzeitig eklatante Ausbildungsdefizite bei der Vermittlung dieser Kompetenzen im Hinblick auf ihre Relevanz im ärztlichen Alltag beklagt (33, 34). In einer Befragung von Jungbauer (35) wurden von den Absolventen praktische Fertigkeiten, Umgang mit dem Patienten, psychosoziale Kompetenz, Kommunikationsfähigkeit und Teamwork in dieser Reihenfolge als die Kompetenzen mit der größten Differenz zwischen dem Stellenwert im Beruf und der Vorbereitung durch das Studium benannt.

In einem Vorschlag für ein Kerncurriculum für die medizinische Ausbildung hat die Bundesvertretung der Medizinstudierenden in Deutschland neun Kompetenzbereiche definiert, von denen zwei explizit kommunikative und soziale Kompetenzen benennen (36). Für den deutschsprachigen Raum existiert mit dem Baseler Consensus Statement eine Experten-Empfehlung für kommunikative und soziale Kompetenzen im Medizinstudium (36, 37). Die Charité schließt mit der Entwicklung von Ausbildungszielen durch die Curriculum-Komitees des Reform- und Regelstudiengangs an die oben skizzierten Entwicklungen in anderen Ländern an: In diesem Ausbildungszielkatalog wurden neun Kompetenzen definiert, von denen eine „Kommunikation, Interaktion und Teamarbeit“ ist.

Aufgrund der Novellierung der Approbationsordnung (38) wurden an vielen Fakultäten verstärkt Anstrengungen unternommen, um die Ausbildung praxisnäher zu gestalten. Mittlerweile haben verschiedene Fakultäten Lehrveranstaltungen zur ärztlichen Gesprächsführung in ihre Curricula integriert, die jedoch hinsichtlich Inhalt, Umfang und didaktischem Konzept sehr heterogen sind: z. B. Regel- und Reformstudiengang in Berlin, Bochum, Dresden, Erlangen, Frankfurt a. M., Göttingen, Greifswald, Hamburg, Heidelberg, Köln, Leipzig, München, Münster, Ulm oder Witten-Herdecke (39-50).

### 1.2.3 Die Übung „Interaktion“ im Reformstudiengang Medizin

Am konsequentesten wurden die Ergebnisse der Forschung bei der Implementierung des Reformstudiengangs Medizin (RSM) an der Charité-Universitätsmedizin Berlin umgesetzt. Der RSM wird seit dem Wintersemester 1999/2000 parallel zum Regelstudiengang angeboten und bildet 10% der Studierenden der Medizin in Berlin aus. Die klassische Aufteilung in Vorklinik und Klinik ist aufgehoben, grundlagenmedizinische und klinische Inhalte sind vom ersten Semester an ineinander verschränkt. Die Ausbildung findet fächerübergreifend und problemorientiert statt, die Semester sind in Themenblöcken organisiert (51). Die Ausbildung im Bereich kommunikative Kompetenzen hat im Vergleich zu traditionellen Studiengängen ein großes Gewicht: In der „Übung Interaktion“ trainieren die Studierenden des RSM in Kleingruppen von sieben Studierenden über zehn Semester kontinuierlich ihre kommunikativen Kompetenzen. Die Übung Interaktion ist eine Pflichtveranstaltung, die zweiwöchentlich im Umfang von drei Zeitstunden stattfindet.

Vermittelt werden in den ersten beiden Semestern Grundlagen der Kommunikation und soziale Kompetenzen. Darauf aufbauend sind ab dem 3. Semester ärztliche Gespräche in den verschiedensten Kontexten und mit verschiedenen Personengruppen sowie in unterschiedlichen Situationen Thema. Dazu gehören beispielsweise Aufklärungs- und Beratungsgespräche, die Kommunikation mit psychiatrischen Patienten oder das Überbringen schlechter Nachrichten. Das vertikale Curriculum ist als Übersicht in Tabelle 1 dargestellt.

**Tabelle 1: Curriculum der Übung „Interaktion“ im RSM - Kurzübersicht**

| Semester | Inhalte |
|---|---|
| **1** | In Kontakt kommen; Soziale Kompetenzen; Selbst-und Fremdwahrnehmung; Verbale und nonverbale Kommunikation, Unterstützung der Zusammenarbeit der POL-Gruppen |
| **2** | Unterstützung der Studierenden im Praxistag[2], Grundlegende Kommunikationsmodelle; Fragen stellen; Gesprächseinstieg und -abschluss; Unterstützung der Zusammenarbeit der POL-Gruppen; |
| **3** | *Anamnese I:* Informationen bekommen: Anamnesen erheben mit dem Schwerpunkt Sexualanamnese |
| **4** | *Anamnese II:* Anamnese erheben mit Fokus auf psychosozialen Aspekten, unterstützend dazu: Video-Analyse des eigenen Gesprächsverhaltens |
| **5** | *Anamnese III:* Anamnesegespräche mit dem Schwerpunkt psychosomatische Patienten |
| **6** | Unterstützung der Studierenden im Blockpraktikum; Informationen geben und gemeinsame Entscheidungen ermöglichen: Aufklärungs- und Beratungsgespräche |
| **7** | Unterstützung der Studierenden im Blockpraktikum; Kommunikation mit Kindern, Jugendlichen und deren Eltern; motivierende Gesprächsführung bei chronisch kranken Patienten |
| **8** | Unterstützung der Studierenden in den Blockpraktika; Kommunikation mit neurologischen und psychiatrischen Patienten sowie psychopathologische Befunderhebung |
| **9** | Unterstützung der Studierenden in den Blockpraktika; Überbringen schlechter Nachrichten; Umgang mit sterbenden /schwerkranken Patienten und deren Angehörigen |
| **10** | Teamarbeit und Fehlermanagement |

Grundlage für die Ausbildung sind die Calgary-Cambridge Observation Guides (CCOG) in einer übersetzten Fassung. Die beiden Guides beschreiben Verhalten, das in der Forschung als hilfreich für die Arzt-Patienten-Kommunikation in Anamnese- sowie Aufklärungs- und Beratungsgesprächen identifiziert worden ist. Sie bilden einen evidenzbasierten, strukturierten Rahmen für die Ausbildung kommunikativer Kompetenzen (1, 52).

Darüberhinaus erlernen die Studierenden in jedem Semester spezifische Inhalte und Techniken. Die Semesterthemen sind abgestimmt auf die medizinischen Inhalte des Curriculums; der Schwierigkeitsgrad und die Komplexität der simulierten Patientengeschichten nehmen mit jedem Semester zu. Ziel der Übung „Interaktion“ ist es, die Studierenden zu befähigen, einen patienten-

[2] Am „Praxistag“ hospitieren die Studierenden des RSM einmal wöchentlich in einer ärztlichen Praxis. Dadurch werden ein früher Patientenkontakt und Einblicke in die ambulante Gesundheitsversorgung gewährleistet. Ziel ist u.a. die Vermittlung von Fertigkeiten zum Aufbau einer tragfähigen Patienten-Arzt-Beziehung auf dem Boden einer patientengerechten Gesprächsführung, geprägt von einer biopsychosozialen Grundhaltung.

zentrierten Gesprächsstil und einen partnerschaftlichen Umgang mit Kollegen zu entwickeln. Die Übung „Interaktion“ ist auch ein Forum, um praktische Erfahrungen und eigenes Verhalten zu reflektieren (53, 54).

Die Lehrveranstaltungen haben einen hohen Übungsanteil in Form von Rollenspielen und Gesprächen mit Simulationspatienten (SP), die ab dem 3. Semester zum Einsatz kommen. In der Regel führt jeder Studierende ein Gespräch pro Semester, zu dem er ein strukturiertes Feedback vom SP, der Gruppe und dem Dozenten erhält. Die Gespräche mit SP werden auf der Basis des CCOG ausgewertet; somit wird Rückmeldung über tatsächlich wirksames Verhalten gegeben. Die Charité unterhält seit der Einführung des Reformstudiengangs Medizin ein Simulationspatientenprogramm, das für die Rekrutierung, das Training und den Einsatz von SP in Lehre und Prüfung zuständig ist.

Dozenten der Übung „Interaktion“ sind Ärzte und Psychologen, die sich speziell für den Unterricht in Gesprächsführung weiterqualifiziert haben: Voraussetzung für eine Dozenten-Tätigkeit ist die Teilnahme an einem vierstündigen Teacher-Training, in dem Informationen zum Konzept der Lehrveranstaltung und zur Arbeit mit Simulationspatienten in der Lehre vermittelt werden. Das Auswerten von Gesprächen mittels CCOG und konstruktivem Feedback sowie die Moderation von Kleingruppen werden aktiv geübt. Desweiteren hospitieren angehende Dozenten mehrmals bei bereits erfahrenen Dozenten in der Lehrveranstaltung. Darüberhinaus werden regelmäßig Fortbildungen zu verschiedenen Themen der Gesprächsführung, die in der Übung „Interaktion“ vermittelt werden, angeboten. Die Dozenten erhalten nach jedem Semester die Ergebnisse der studentischen Evaluation als Feedback.

## 1.3 Überprüfung praktischer Fertigkeiten in der medizinischen Ausbildung

### 1.3.1 Argumente für die Prüfung kommunikativer Kompetenzen

In der Forschung wird die regelmäßige Überprüfung der kommunikativen Kompetenzen während der Ausbildung empfohlen. Verschiedene Gründe sprechen dafür, kommunikative Kompetenzen nicht nur zu lehren, sondern auch zu prüfen (55-58):

1) Studierende tendieren dazu, Ausbildungsziele, die nicht geprüft werden, weniger wichtig zu nehmen. Aus der Perspektive von Studierenden wird Erfolg im Studium mit dem Bestehen der Prüfungen gleichgesetzt. Dieser Sachverhalt wird oft mit dem Satz „assessment drives learning“ beschrieben. Dass Prüfungen das Lernverhalten der

Studierenden stark beeinflussen, wird oft unterschätzt; Studierende konzentrieren sich auf die Inhalte, von denen sie wissen oder vermuten, dass diese Gegenstand von Prüfungen sein werden. Themen, die nicht geprüft werden, drohen, vernachlässigt zu werden. Wenn kommunikativen Kompetenzen der Stellenwert zugeschrieben werden soll, der ihnen als Pflichtfach über zehn Semester zugedacht ist, sollten sie daher auch geprüft werden.

2) Durch eine Leistungsbeurteilung im Rahmen einer Prüfung erhalten die Studierenden eine Rückmeldung über ihre kommunikativen Kompetenzen und ihren Lernerfolg.

3) Durch eine (bestehensrelevante) Prüfung wird sichergestellt, dass die Absolventen in diesem Bereich kompetent sind. Die Universität sichert damit die Qualität der Ausbildung; für die Absolventen ist dies angesichts der in 1.1 dargestellten Relevanz gelungener Arzt-Patienten-Kommunikation wichtig.

4) Für die Beurteilung kommunikativer Kompetenzen sind Prüfer notwendig, die die Leistungen der Studierenden bewerten. Die Aufnahme dieser Inhalte in den Prüfungskanon kann im Sinne einer Fakultätsentwicklung zur Reflexion der Prüfer über die eigene Arzt-Patienten-Kommunikation beitragen. Im besten Falle wird sowohl die Lehre als auch die Patientenversorgung dadurch positiv beeinflusst.

5) Prüfungen sichern die Qualität des Curriculums: die Prüfungsergebnisse dienen dem Feedback an Lehrende und Planende, um das Curriculum zu optimieren.

### 1.3.2 Methoden zur Prüfung praktischer Fertigkeiten

Es stellt sich die Frage, welches Prüfungsformat zur Beurteilung praktischer Fertigkeiten geeignet ist. Die „Prüfung am Krankenbett“ als traditionelle Form der Überprüfung klinisch-praktischer Kompetenzen weist mehrere limitierende Faktoren auf, die die Qualität der Prüfung erheblich mindern: Ein Studierender wird dabei von sehr wenigen, im Extremfall nur einem Prüfer beurteilt. Er kann seine Fertigkeiten und sein Wissen nur an einem oder zwei Patienten demonstrieren. Unterschiede zwischen den Prüfern, den Prüfungsinhalten und den Patienten führen zu einer sehr ungenauen Leistungsbeurteilung. Die Stichprobe des geprüften Wissens ist zu klein, so dass das Prüfungsergebnis im Hinblick auf das zu erwartende Wissen nicht verallgemeinerbar ist. In Nordamerika führten die inakzeptabel geringen Übereinstimmungen zwischen Prüferurteilen in traditionellen klinisch-mündlichen Prüfungen Ende der 60er Jahre zur Suche nach anderen Prüfungsformaten (59).

Zur Kategorisierung von Prüfungsinhalten hat Miller (60) ein Pyramiden-Modell entwickelt, in dem die Kompetenzentwicklung von Studierenden als aufeinander aufbauende Stufen dargestellt ist. Die Basis der Pyramide repräsentiert das faktische Wissen („knows"), gefolgt vom prozeduralen oder angewandten Wissen („knows how"). Die Fähigkeit zur Umsetzung im Handeln („shows how") liegt darüber, mit der tatsächlichen Performanz im ärztlichen Alltag („does") an der Spitze. Jedes Wissen auf einer höheren Ebene setzt auch Wissen der darunterliegenden Stufen voraus.

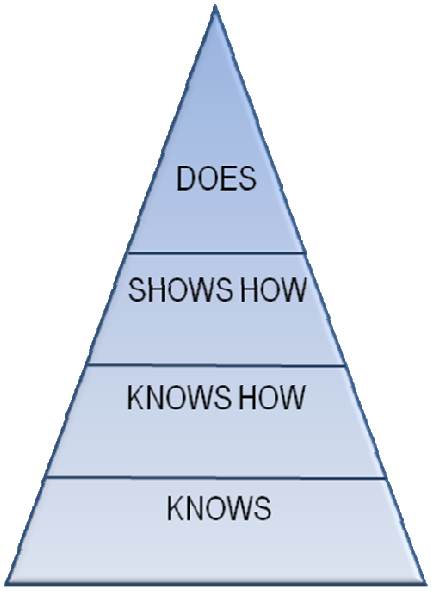


**Abbildung 1: Pyramidenmodell der Kompetenzentwicklung nach Miller (1990)**

Kein Prüfungsformat ist geeignet, um alle Facetten gleichzeitig zu prüfen. Während sich für die beiden unteren Stufen schriftliche Formate wie z. B. Multiple-Choice-Fragen gut eignen, sind diese nicht adäquat, um die Umsetzung im Handeln zu überprüfen. Dazu wäre es notwendig, zu beobachten, wie sich die Prüflinge in einem echten Setting (z. B. auf der Station) verhalten. Verfahren, um praktische Fertigkeiten auf der „does"-Ebene zu beurteilen, sind das Mini-Clinical Evaluation Exercise (Mini-Cex), bei dem der Prüfling über einen längeren Zeitraum in verschiedenen realen Situationen von verschiedenen Prüfern beobachtet und beurteilt wird sowie die 360-Grad-Beurteilung, bei der mehrere Beteiligte die Performanz des Prüflings aus verschiedenen Perspektiven (z. B. ärztliche Kollegen, Pflegekräfte, Patienten) beurteilen (61). Diese Verfahren sind jedoch sehr aufwendig in der Umsetzung und eher in höheren Ausbildungsstufen bzw. der Weiterbildung sinnvoll einsetzbar. Eine weitere Möglichkeit, praktische Fertigkeiten valide zu prüfen, besteht in der Simulation realer Situationen; dies

entspricht der Stufe „shows how“ in der Miller’schen Pyramide. In der medizinischen Ausbildung hat sich ein spezielles Format zur Prüfung klinisch-praktischer Fertigkeiten mittels Simulation etabliert, das für die Prüfung kommunikativer Kompetenzen hervorragend geeignet ist: die „objective structured clinical examination“ (OSCE).

### 1.3.3 Die Prüfungsmethode “Objective Structured Clinical Examination” (OSCE)

Der OSCE als Prüfungsformat wurde in den siebziger Jahren von Harden (62) entwickelt, um die Mängel der traditionellen Prüfung klinisch-praktischer Kompetenzen zu überwinden. Ein OSCE ist ein Prüfungsparcours, der von allen Studierenden im Rotationsverfahren durchlaufen wird. Der Parcours beinhaltet verschiedene Aufgaben (sogenannte „Stationen“), die in einer vorgegebenen Zeitspanne bewältigt werden müssen und deren Schwierigkeitsgrad an den Ausbildungsstand angepaßt ist. Im OSCE kann ein großes Spektrum praktischer Fertigkeiten geprüft werden von Anamnese-, Aufklärungs-, Beratungs- und/oder Untersuchungsstationen über weitere Aufgaben wie z. B. das Schreiben von EKGs, Blutentnahme am Modell, Durchführung von Notfallmaßnahmen, Interpretation von Röntgenbildern/CT-Bildern oder Mikroskopieren. Je nach Aufgabenstellung ist der Einsatz von Modellen und/oder Simulationspatienten sinnvoll, um eine quasi-reale Situation zu schaffen; dabei ist die Standardisierung besonders wichtig, um die stets gleiche, wiederholte Darstellung eines Patientenfalls durch den SP und damit gleiche Bedingungen für alle Prüflinge zu gewährleisten. Die Interaktion des Prüflings findet in diesem Fall vorrangig mit dem Simulationspatienten statt, nicht mit dem Prüfer.

Alle Stationen dauern gleich lang, in der Regel 5-20 Minuten in Abhängigkeit von der Komplexität der Aufgabenstellungen. Alle Prüflinge rotieren durch alle Stationen. An jeder Station wird die Leistung der Studierenden von einem Prüfer auf einem standardisierten Beurteilungsbogen dokumentiert.

### 1.3.4 Einordnung der Prüfungsmethode OSCE aus testtheoretischer Perspektive

Im Gegensatz zu Prüfungen am Krankenbett erlaubt ein OSCE eine detaillierte statistische Auswertung zur Qualitätssicherung und zur Evaluation des Curriculums. Für die Beurteilung der Qualität einer diagnostischen Methode sind die sogenannten Hauptgütekriterien von zentraler Bedeutung: Objektivität, Reliabilität und Validität.

Die Objektivität eines Tests bezeichnet das Ausmaß, in dem die Ergebnisse eines Tests unabhängig vom Testleiter (hier: Prüfer) sind. Es wird zwischen Durchführungs-, Auswertungs- und Interpretationsobjektivität unterschieden.

Die Reliabilität (Zuverlässigkeit) beschreibt die Messgenauigkeit, mit der ein Testverfahren eine Merkmalsdimension erfaßt. Dabei bleibt die inhaltliche Treffsicherheit außen vor. Zur Bestimmung der Zuverlässigkeit eines Tests gibt es verschiedenen Methoden. Die Genauigkeit einer Messung ist umso relevanter, je größer die Konsequenzen des Ergebnisses sind.

Die Validität (Gültigkeit) gibt an, wie genau der Test das Merkmal misst, das er erfassen soll. Bei hoher Validität können die Ergebnisse generalisiert werden, d. h. von dem Verhalten in der Testsituation kann auf Verhalten außerhalb der Testsituation geschlossen werden (63). Unterschieden werden die inhaltliche Validität, bei der die Aufgaben eines Tests Stichproben aus dem zu erfassenden Zielmerkmal darstellen, sowie die kriteriumsbezogene Validität und die Konstruktvalidität. Die Kriteriumsvalidität wird ermittelt als Korrelation zwischen Testergebnis und interessierendem Merkmal (Kriterium), wobei je nach Zeitpunkt der Erhebung der Kriteriumswerte von konkurrenter oder prädiktiver Validität gesprochen wird. Die Konstruktvalidität wird durch das Testen von Hypothesen über das Konstrukt und seine Relationen zu anderen Variablen untersucht. Messen verschiedene Methoden das Konstrukt übereinstimmend, spricht man von konvergenter Validität. Die diskriminante Validität zielt im Gegensatz dazu auf die Unterscheidung des Zielkonstrukts von anderen Konstrukten ab (64).

Im Folgenden werden die drei Hauptgütekriterien in Bezug auf das Prüfungsformat OSCE diskutiert.

#### *1.3.4.1 Objektivität*

Im Unterschied zu herkömmlichen Verfahren zur Prüfung praktischer Fertigkeiten zeichnet sich der OSCE durch eine hohe Durchführungs-, Auswertungs- und Interpretationsobjektivität aus, da der Ablauf stark standardisiert ist: Die Aufgabenstellungen und Beurteilungsbögen sind für alle Studierenden identisch, die Darstellung durch Simulationspatienten ist möglichst standardisiert und die soziale Interaktion zwischen Prüfer und Prüfling auf ein Minimum beschränkt. Je weniger Einfluss die Person des Prüfers auf die Durchführung der Prüfung hat, desto höher ist die Durchführungsobjektivität.

Als Beurteilungsinstrumente werden zumeist Checklisten eingesetzt, deren Inhalt vorher festgelegt wurde und den Prüfern bekannt ist. Die Prüfer sind im Gegensatz zu mündlichen Prüfungen in einer beobachtenden Rolle; sie beurteilen und dokumentieren die Leistungen der Studierenden. Um eine hohe Objektivität und Testfairness zu erreichen, ist es wichtig, für alle Prüfer gültige Verhaltensregeln ähnlich denen eines Testleiters zu etablieren (wie z. B. keine eigenen Fragen zur Aufgabe zu stellen oder die Leistung des Prüflings nicht zu kommentieren) und klare Performanzkriterien zu definieren (57).

#### *1.3.4.2 Reliabilität*

Die Inter-Rater-Reliabilität bezeichnet die Übereinstimmung zwischen Urteilen verschiedener Prüfer, die Intra-Rater-Reliabilität die Übereinstimmung der Urteile eines Prüfers zu zwei Zeitpunkten. Beide sind durch das standardisierte Bewertungssystem in der Regel hoch (59). Hervorzuheben ist dagegen die mögliche Einschränkung der Reliabilität durch die sogenannte Kontextspezifität (auch Inter-Case-Reliabilität genannt): Die Performanz von Prüflingen über verschiedene Fälle ist nicht konsistent; die Leistung eines Prüflings in einer Station ist ein schlechter Prädiktor für die Bewältigung der nächsten Aufgabe. Um die Inter-Case-Reliabilität zu erhöhen, ist daher eine große Anzahl von Stationen mit verschiedenen Problemen (Aufgaben, Inhalten) in einer OSCE-Prüfung notwendig, was zu sehr langen Testzeiten von mehreren Stunden führt (65).

Die Genauigkeit, mit der ein Test ein Merkmal erfaßt, läßt sich mittels verschiedener Verfahren bestimmen. Bei der Berechnung der Reliabilität von OSCE-Prüfungen kommen Konsistenzanalysen im Sinne der klassischen Testtheorie zum Einsatz, zunehmend wird auch die Generalisierbarkeitstheorie als Erweiterung der klassischen Testtheorie herangezogen. Der Ansatz der Generalisierbarkeitstheorie erlaubt es, die Varianzanteile der Beurteiler und der Situation gleichzeitig in einer Analyse zu berücksichtigen (66).

#### *1.3.4.3 Validität*

Die Gültigkeit eines Tests drückt aus, ob ein Test das Merkmal mißt, das er messen soll. Nur bei hoher Validität ist die Generalisierung anhand der Testergebnisse auf andere Situationen zulässig (63). Um eine hohe inhaltliche Validität in einer OSCE-Prüfung zu erzielen, sollte ihr ein sorgfältig konstruierter Blueprint zugrunde liegen. Diese Blaupause ist eine Matrix, in der Prüfungs-

aufgaben den Ausbildungszielen und den zu testenden Kompetenzen zugeordnet werden. Damit wird sichergestellt, dass die Aufgaben Stichproben des zu erfassenden Zielmerkmals darstellen und eine breit gestreute Verhaltensstichprobe des Universums des erwarteten Verhaltens abbilden (sampling), die einen Repräsentationsschluss auf die Kompetenz der Prüflinge erlaubt (59). Bei der Konstruktion der Prüfung muss abgewogen werden zwischen möglichst vielen Stationen zur Reliabilitätssteigerung und der Gefahr, dass komplexe Fertigkeiten durch eine kurze Stationslänge fragmentiert werden, was wiederum die Validität einschränkt (67). Bei der Konzeptionierung der Aufgaben ist große Sorgfalt im Sinne möglichst realitätsnaher Ausgestaltung Voraussetzung für eine valide Prüfung; wenn Zeitbegrenzung und Setting zu unrealistischen Stationen führen, werden anderenfalls die Schauspielfähigkeiten der Prüflinge getestet. In Studien zur Validität von OSCEs werden meistens Gruppenunterschiede geprüft oder Korrelationsmaße berechnet; entweder als Vergleich der Testergebnisse von Studierenden verschiedener Ausbildungsstufen (Konstruktvalidität) oder als Vergleich mit anderen Testergebnissen (Kriteriumsbezogene Validität).

### 1.3.5 Stellenwert des OSCE in der medizinischen Ausbildung

Der OSCE ist mittlerweile ein international etabliertes und stark beforschtes Prüfungsformat mit verschiedenen Weiterentwicklungen. Simulationen zur Prüfung klinisch-praktischer Fertigkeiten sind Standard in Nordamerika (67). In Kanada und den USA sind OSCEs seit den neunziger Jahren Teil der Zulassungs- bzw. Staatsprüfungen mit bis zu 16.000 Prüflingen pro Jahr.

Im deutschsprachigen Raum setzt sich das Verfahren im Zuge der Studienreformen erst langsam durch. Vorreiter war der Reformstudiengang Medizin an der Charité in Berlin, wo seit 1999 bestehensrelevante OSCE-Prüfungen in acht von zehn Semestern durchgeführt werden. Die Novellierung der Approbationsordnung für Ärzte im Jahr 2002 mit einer größeren Betonung der praxisnahen Ausbildung dürfte das Interesse an dieser Prüfungsmethode steigern und zu einer größeren Verbreitung in Deutschland führen. Bei einer Befragung im Jahr 2005 gaben nur sieben der 36 medizinischen Fakultäten in Deutschland an, OSCE-Prüfungen unter Beteiligung von SP durchzuführen (68). In einer Absolventenbefragung bewerteten Facharztprüflinge Prüfungen im OSCE-Format als optimale Vorbereitung für die spätere ärztliche Tätigkeit (33).

### 1.3.6 OSCE-Prüfungen im Reformstudiengang Medizin

Aufgrund § 41 der Ärztlichen Approbationsordnung entfällt für die Studierenden des RSM der erste Abschnitt der ärztlichen Prüfungen („Modellstudiengangsparagraph“) (38). Stattdessen werden die Studierenden des Reformstudiengangs am Ende jedes Semesters bestehensrelevant fakultätsintern geprüft, wobei großer Wert auf innovative Prüfungsformen gelegt wird.

In den Semesterabschlussprüfungen werden kognitives Wissen und klinisch-praktische Fertigkeiten der Studierenden bestehensrelevant überprüft; dabei kommen Multiple Choice- und OSCE-Prüfungen zum Einsatz. Ein OSCE besteht im RSM in der Regel aus 5 bis 10 Stationen. Die Dauer der Stationen ist abhängig von der Komplexität der Aufgabenstellungen und liegt in den ersten Semestern bei fünf Minuten, in den höheren Semestern bei längstens 10 Minuten. Insgesamt werden die Studierenden im Laufe ihres Studiums ungefähr acht Stunden geprüft im Hinblick auf ihre klinisch-praktischen Fertigkeiten (69).

Die Aufgabenstellungen orientieren sich an den Lernzielen der Themenblöcke in einem Semester. Je nach Aufgabe kommen Modelle oder SP zum Einsatz. Die Prüfer füllen für jeden Studierenden an jeder Station eine Checkliste mit drei Antwortmöglichkeiten (erfüllt – teilweise erfüllt – nicht erfüllt) aus; die Zahl der Items variiert je nach Aufgabenstellung. Die erreichten Punkte in jeder Station werden summiert und in Prozentwerte umgerechnet. Das Gesamtergebnis jedes Prüflings errechnet sich aus dem arithmetischen Mittel der Prozentwerte aller Stationen. Die Bestehensgrenze wurde kriteriumsorientiert auf 60% festgelegt. Die Studierenden erhalten ein Feedback in Form eines Ergebnisbogens, auf dem das Abschneiden in jeder Station und der Vergleich mit der Prüfungskohorte dargestellt sind. Die Inhalte der Checkliste sind den Studierenden nicht bekannt.

Kommunikative Kompetenzen wurden bisher nicht explizit im OSCE geprüft. Einzelne Items in den Checklisten beziehen sich zwar in einigen Stationen auf Elemente der Gesprächsführung wie beispielsweise Vorstellung und Begrüßung, diese sind jedoch nicht standardisiert. Andere, in der Übung Interaktion vermittelte Inhalte sind nicht Gegenstand der Prüfung. Daraus resultiert bei den Prüfern immer wieder Unbehagen, da im bisherigen Bewertungssystem Studierende die gleiche Punktzahl erhalten, solange sie die in der Checkliste enthaltenen Inhalte ansprechen, unabhängig davon, wie sie auftreten und wie sie mit den Simulationspatienten kommunizieren. Von den Prüfern wahrgenommene Kompetenzunterschiede in der Gesprächsführung werden nicht abgebildet.

## 1.4 Überprüfung kommunikativer Kompetenzen mittels OSCE

Für die Überprüfung kommunikativer Kompetenzen läßt sich eine konsistente Darstellung nur in Prüfungsformaten mit Simulationspatienten erreichen (Boursicot & Roberts, 2005). In Situationen mit Simulationspatienten können viele verschiedene ärztliche Gesprächssituationen in standardisierter Form (mit-)geprüft werden. Aufgrund seiner hervorragenden Eignung als Prüfungsmethode für die Prüfung kommunikativer Kompetenzen hat sich der OSCE in vielen Ländern durchgesetzt: So berichten 11 der 13 britischen Universitäten, die kommunikative Kompetenzen prüfen, dass sie OSCE als Prüfungsmethode einsetzen (16). Im amerikanischen Examen USMLE Step 2 Clinical Skills (United States Medical Licensing Examination), das die Voraussetzung für die amerikanische Lizenz ist, werden seit 2004 kommunikative Kompetenzen im Rahmen eines OSCEs geprüft; ausländische Absolventen müssen ihre kommunikativen Fertigkeiten schon seit 1998 in dieser Form demonstrieren (Educational Commission for Foreign Medical Graduates, ECFMG). In Kanada werden seit 1993 im OSCE als Bestandteil des Staatsexamens (QE II) auch kommunikative Kompetenzen geprüft. Während die Beurteilung der Performanz in den USA häufig auch durch SP stattfindet, verläßt man sich in Kanada ebenso wie in Europa lieber auf die Beurteilung durch akademische Mitarbeiter der Fakultät (70).

## 1.5 Forschungsstand zur Prüfung von kommunikativen Kompetenzen

Als Beurteilungsinstrumente werden in OSCE-Prüfungen sowohl Checklisten als auch globale Ratingskalen eingesetzt. Checklisten enthalten einzelne, zumeist zahlreiche Items, die detailliert Verhalten beschreiben, dessen Auftreten bzw. Ausbleiben dokumentiert wird. Ratingskalen bestehen meist aus weniger Items, die Verhalten auf einem globaleren Niveau beschreiben, und geben ein mehrfach gestuftes Merkmalskontinuum vor. Beide Formate sind mit Vor- und Nachteilen verbunden, die mittlerweile intensiv beforscht worden sind: Detaillierte Checklisten führen meist zu guter Reliabilität der Prüfung, sind oftmals allerdings auch trivial und nicht immer geeignet, Performanz adäquat abzubilden. Zudem kann abweichendes Verhalten, das auf der Checkliste nicht aufgeführt ist, weder honoriert noch sanktioniert werden. Mit Checklisten wird weniger Expertise belohnt, die sich in diesem Format schlecht abbilden läßt, sondern Gründlichkeit im Sinne eines mechanischen Abhakens von Schritten, das Novizen kennzeichnet. Die Expertise von Prüfern wird beim Einsatz detaillierter Checklisten minimiert (59, 65, 71).

Während Checklisten sich besonders zur Prüfung von Stufenschemata (z. B. Reanimation) oder Untersuchungstechniken eignen, haben sich globale Beurteilungsinstrumente bewährt, um

komplexe Fertigkeiten valide zu messen. In Studien zeigten sich globale Ratings bei der Beurteilung der Performanz auf einem fortgeschrittenen Ausbildungslevel sowie der Erfassung von Nuancen überlegen (71, 72). Dabei haben sich globale, fünf- bis siebenstufige Ratings, die Experten von genutzt wurden, als mindestens ebenso reliabel wie Checklisten im OSCE erwiesen (73). Die Verwendung holistischer Skalen erfordert jedoch ein intensives Rater-Training und die Definition klarer Kriterien, um die Subjektivität der Prüferurteile reduzieren (57). Über die Ausgestaltung von Trainingsmaßnahmen wird in Publikationen jedoch häufig nur mangelhaft berichtet, z. B. durch die alleinige Angabe der Trainingsdauer (74).

In der Literatur wird für die Prüfung kommunikativer Kompetenzen im OSCE der Einsatz von globalen Beurteilungsinstrumenten als vorteilhafter im Vergleich zu Checklisten angesehen (59, 67). Die Beurteilung der kommunikativen Kompetenzen bei Medizinstudierenden sollte im Sinne hoher ökologischer Validität nicht losgelöst vom ärztlichen Kontext stattfinden. Fachwissen ist zu einem gewissen Grad eine Voraussetzung für eine gute ärztliche Gesprächsführung, da kommunikative Kompetenzen keine generellen Fähigkeiten sind, die losgelöst vom Inhalt erfaßt werden könnten. Die Performanz ist nicht völlig unabhängig vom Inhalt der Aufgabe; der Aspekt der Kontextspezifität ist daher auch bei der Messung kommunikativer Kompetenzen zu berücksichtigen.

## 1.6 Synopsis und Zielsetzung

Angesichts der vielfältig nachgewiesenen positiven Effekte gelungener Arzt-Patienten-Kommunikation wird die Notwendigkeit der Vermittlung kommunikativer Kompetenzen während des Studiums nicht mehr infrage gestellt. Dem Training kommunikativer Kompetenzen in der medizinischen Ausbildung kommt zunehmend Bedeutung zu. Als Konsequenz der Forschungsergebnisse wurden an vielen medizinischen Fakultäten Kommunikationstrainings in die Curricula implementiert und Simulationspatienten in die Lehre integriert; beides ist heute im englischsprachigen Raum Standard in der medizinischen Ausbildung. Deutschland hinkt dieser Entwicklung trotz zunehmender Aktivität in diesem Bereich hinterher. Im Reformstudiengang Medizin in Berlin werden kommunikative Kompetenzen mit einem in Deutschland einzigartigen Konzept schon seit 1999 kontinuierlich während des Studiums vermittelt.

Wie in 1.3.1 dargelegt, sprechen viele Gründe dafür, ärztliche Gesprächsführung nicht nur zu lehren, sondern den Kompetenzerwerb auch zu überprüfen. Die Prüfung der kommunikativen Kompetenzen wird daher in der Forschung als integraler Bestandteil der medizinischen Aus-

bildung angesehen; international wird der Prüfung kommunikativer Kompetenzen mittlerweile große Bedeutung zugemessen. Eine Prüfungsmethode, die sich dafür besonders eignet, ist der OSCE. OSCE-Prüfungen finden im Reformstudiengang Medizin in Berlin nach fast jedem Semester statt; kommunikative Kompetenzen werden dabei bisher nicht (systematisch) geprüft. Aus den in 1.3.1 genannten Gründen gibt es schon länger Überlegungen, den Kompetenzerwerb der Studierenden im Reformstudiengang zu überprüfen. Dies ließe sich in den OSCE-Prüfungen annähernd kostenneutral und mit den vorhandenen Ressourcen durchführen, wenn die Prüfer (oder die SPs) zusätzlich zu den klinisch-praktischen Fertigkeiten auch die kommunikativen Kompetenzen beurteilen würden. Die Implementation im Rahmen des OSCE entspricht dem Stand der Forschung.

Im englischsprachigen Raum ist in den letzten zehn Jahren eine Fülle von Instrumenten zur Beurteilung kommunikativer Kompetenzen veröffentlicht worden, sowohl Checklisten als auch globale Ratingskalen. Während die Reliabilität (meist als Inter-Rater-Reliabilität) zum Teil untersucht worden ist, stehen Untersuchungen zur Validität bei den meisten Instrumenten aus. Insbesondere der Vergleich mit anderen Instrumenten im Sinne der Konstruktvalidität ist selten; dort, wo Vergleiche durchgeführt worden sind, waren die Korrelationen meist relativ niedrig (75). Dagegen existiert meines Wissens bisher kein einziges deutschsprachiges Instrument zur Prüfung kommunikativer Kompetenzen, dessen psychometrische Qualitäten untersucht worden sind.

Ziel dieser Untersuchung ist es, diese Lücke zu schließen. Vor dem Hintergrund einer Fülle von englischsprachigen Instrumenten scheint eine Adaption und Überprüfung eines vorhandenen Instruments sinnvoller als eine weitere Neuentwicklung. Ziel der Arbeit ist die Identifikation und Validierung eines geeigneten Beurteilungsinstruments, das für den deutschsprachigen Raum adaptiert und im Rahmen von OSCE-Prüfungen eingesetzt werden kann. Das Beurteilungsinstrument sollte idealerweise bereits im Hinblick auf seine psychometrische Qualität untersucht worden sein. Mittels verschiedener Validierungsstrategien betreffend Kriteriums- und Konstruktvalidität soll in dieser Arbeit geprüft werden, inwieweit das Instrument für die Prüfung kommunikativer Kompetenzen bei Medizinstudierenden im Rahmen von OSCE-Prüfungen geeignet ist.

Kriteriumsvalidität liegt dann vor, wenn das Ergebnis eines Tests zur Messung eines Konstrukts mit Messungen eines korrespondierenden Kriteriums übereinstimmt (64). Die Kriteriumsvalidierung ist in ihrem Anwendungsbereich oftmals stark eingeschränkt, da nicht immer ein adäquates Außenkriterium herangezogen werden kann. Mangels externer Kriterien, gegen

welche die Validität eines Instruments als „Goldstandard“ getestet werden kann, wird als Strategie zur Validierung die Beurteilung durch Experten mit hoher Kompetenz im interessierenden Bereich empfohlen (76, 77). Das Urteil der Experten wird dann als Referenzkriterium definiert.

Im interessierenden Einsatzfeld eines Beurteilungsinstruments im Rahmen eines OSCEs sind die späteren Prüfer vor allem Experten ihres jeweiligen medizinischen Fachgebiets, jedoch nicht unbedingt ausgewiesene Experten ärztlicher Gesprächsführung. Es muss daher geklärt werden, inwieweit die Urteile von Prüfern nach Trainingsmaßnahmen mit den Urteilen von Kommunikations-Experten übereinstimmen, deren zusätzliche regelmäßige Anwesenheit in der Prüfung unrealistisch ist. Weiterhin interessieren die Urteile der Simulationspatienten als Gesprächspartner der Prüflinge. Sie sind regelmäßig anwesend in der Prüfung, und ihr Urteil könnte in Bewertungen einfließen. Möglicherweise unterscheiden sich die Bewertungen aus der Perspektive der Angesprochenen auch von denen der Beobachter (Prüfer und Experten). Da Test- und Kriteriumswerte zum selben Meßzeitpunkt erhoben werden, können damit Aussagen zur konkurrenten Validität getroffen werden.

Im Rahmen der Konstruktvalidierung werden Hypothesen über das Konstrukt und seine Relationen zu anderen Konstrukten abgeleitet und überprüft. Konvergente Validität liegt vor, wenn unterschiedliche Methoden dasselbe Konstrukt übereinstimmend messen (64). Durch den Vergleich mit einem anderen Instrument zur Messung kommunikativer Kompetenzen sollen daher Belege für die Konstruktvalidität gewonnen werden. Um sicherzustellen, dass mit dem neuen Instrument relevante Aspekte der Gesprächsführung, wie sie im Reformstudiengang Medizin vermittelt werden, erfaßt werden, soll das dem Unterricht in „Interaktion“ zugrunde liegende Instrument (CCOG) in Form einer Checkliste in die Studie einbezogen werden. Eine hohe Übereinstimmung zwischen beiden Instrumenten würde dafür sprechen, dass das gleiche Konstrukt - kommunikative Kompetenzen - gemessen wird (konvergente Validität).

Ein weiteres Beurteilungskriterium für die Eignung des Instruments ist die Abschätzung der diskriminanten Validität. Diskriminante Validität liegt dann vor, wenn sich das Zielkonstrukt hypothesenkonform von anderen Konstrukten unterscheidet. Zur Beurteilung der diskriminanten Validität sollen in dieser Arbeit die Prüfungsergebnisse der Studierenden im OSCE herangezogen und untersucht werden, ob das Instrument die kommunikativen Kompetenzen ausreichend unabhängig von den klinischen Fertigkeiten erfassen kann. Dies wäre ein weiterer Beleg für die Konstruktvalidität.

## 1.7 Fragestellungen und Hypothesen

Aus den dargestellten Forschungsergebnissen werden folgende Hypothesen und Fragestellungen zur Kriteriumsvalidität (Nr. 1 und 2) und zur Konstruktvalidität (Nr. 3 und 4) abgeleitet:

Zusammenhangshypothesen

1. Es wird erwartet, dass die Urteile von trainierten Prüfern und Experten als Referenzkriterium auf dem Instrument übereinstimmen (konkurrente Validität).
2. Es wird erwartet, dass die Urteile von trainierten Simulationspatienten und Experten als Referenzkriterium auf dem Instrument übereinstimmen (konkurrente Validität).
3. Es wird erwartet, dass die Meßwerte auf dem Instrument mit Meßwerten, die mittels eines anderen Instruments zur Beurteilung kommunikativer Kompetenzen (CCOG) erhoben werden, übereinstimmen (konvergente Validität).

Offene Fragestellung

4. Liegt eine Konfundierung von kommunikativen Kompetenzen mit klinisch-praktischen Fertigkeiten vor? (diskriminante Validität)

Fragestellung 4 wird als offene Forschungsfrage gestellt, da keine theoretischen Vorannahmen zur Ausprägung einer zu erwartenden Beziehung vorliegen.

# 2 Methoden

## 2.1 Auswahl und Adaption eines Beurteilungsinstruments

Aufgrund der Forschungslage zur Beurteilung kommunikativer Kompetenzen (vgl. 1.5) wurde einem globalen Beurteilungsinstrument der Vorzug gegeben. Das Instrument sollte zum Beurteilen beobachteter Interaktion von Medizinstudierenden mit Simulationspatienten geeignet sein; zur psychometrischen Qualität sollten bereits Daten vorliegen. Im Hinblick auf die Integration in eine OSCE-Prüfung ist der Einsatz eines globalen Ratinginstruments auch unter Praktikabilitätsaspekten im Vergleich zu den meist langen Checklisten vorteilhafter: Während die Prüflinge zur nächsten Station rotieren, haben die Prüfer eine Minute Zeit, um ihre Beurteilung abzugeben; in dieser Zeit eine längere Checkliste auszufüllen, wäre unrealistisch. Das Instrument sollte daher nicht zu viele Items aufweisen.

In den Literatur-Datenbanken Medline, Psyndex, Psychinfo und Social Sciences Citation Index wurde bis zum Jahr 2004 nach einem Instrument mit den Stichwörtern „communication skills", „interpersonal skills", „assessment", „evaluation", „measurement", „validation" „global rating", „OSCE" und „medical education" gesucht. Nach Sichtung aller Treffer erfüllten drei veröffentlichte Instrumente die oben genannten Kriterien:

- die ABIM Rating Scale mit 10 Items (78)
- die Interpersonal Skills Rating Form mit 13 Items (79) sowie
- ein globales Rating aus Toronto mit vier Items (80)

Verschiedene Argumente sprachen für das letzte Instrument: Das von B. Hodges und J. McIlroy beschriebene Instrument wurde in Kooperation mit Cleo Boyd, einer kanadischen Linguistin, sehr sorgfältig entwickelt. Es ist nicht auf eine spezielle Kommunikationssituation zugeschnitten, sondern für die Beurteilung dyadischer Interaktionen zwischen Arzt und Patient in verschiedenen Kontexten geeignet. In der o.g. Studie, in der Studierende in einem OSCE mit zehn Stationen von ärztlichen Prüfern beurteilt wurden, konnten Hodges & McIlroy zeigen, dass Kompetenzunterschiede zwischen Studierenden verschiedener Semester auf der globalen Ratingskala abgebildet werden können und damit einen Hinweis für die Konstruktvalidität erbringen. Das Instrument wird in Kanada in OSCE-Prüfungen verschiedener Fächer an zahlreichen Fakultäten eingesetzt (B. Hodges, persönliche Mitteilung, 07.09.2004). Weiterhin sind die beiden erstgenannten Instrumente aus der Perspektive des (Simulations-)Patienten formuliert, das dritte ist neutral formuliert und so auch für die Prüfer passend. Da es außerdem deutlich kürzer ist als die

beiden anderen, eignet es sich für die Implementation im OSCE am besten und wurde daher ausgewählt. Das globale Rating besteht aus den Dimensionen

- Eingehen auf die Gefühle und Bedürfnisse des Patienten (Empathie)
- Logischer Zusammenhang des Gesprächs (Struktur)
- verbaler Ausdruck und
- nonverbaler Ausdruck,

die jeweils auf einer fünfstufigen Skala bewertet werden. Die Punktwerte werden zu einem Gesamtwert summiert. Für jede Dimension sind Verhaltensanker für eine schwache, mittlere und hervorragende Leistung beschrieben.

Mit Zustimmung von B. Hodges übersetzten die an dem Forschungsprojekt Beteiligten (A. Fröhmel, I. Mühlinghaus, H. Ortwein, S. Scheffer) das Instrument unabhängig voneinander. Daraufhin wurden die vier individuellen Übersetzungen verglichen und diskutiert, bis schließlich eine einheitliche Fassung vorlag. Dieses Vorgehen wurde gewählt, um eine möglichst zuverlässige Übersetzung sicherzustellen. An der übersetzten Fassung wurden einige Modifikationen vorgenommen: Die numerische Skalierung des Bewertungsmaßstabs (5 = excellent, 1 = poor) wurde der deutschen Notenvergabe angepaßt, um Verwechslungen zu vermeiden und eine bessere Passung in den deutschen Kontext zu gewährleisten (1 = beste Leistung, 5 = schlechteste Leistung). Die Beschreibungen mittelmäßiger Performanz (3 Punkte) wurden entfernt, da sie z. T. nicht genau der Mitte entsprachen oder neue Inhalte enthielten. Die Beschreibungen der Skalen-Endpunkte wurden dahingehend angepaßt, dass sich die Beschreibungen einer schlechten und sehr guten Leistung inhaltlich entsprechen, so dass eine bipolare Ratingskala vorliegt. Dies begünstigt Intervallskalenqualität (81). Die übersetzte und modifizierte Fassung des „Berliner Global Ratings“ befindet sich ebenso wie das Original im Anhang (S. 83f).

## 2.2 Notwendigkeit eines Rater-Trainings

Bei der Verwendung von globalen Beurteilungsinstrumenten stellt die individuelle Interpretation von Items und die daraus resultierende mangelnde Inter-Rater-Reliabilität eine Gefahr dar. Die Übereinstimmung zwischen Ratern kann durch Trainingsmaßnahmen erhöht werden und ist bei der Verwendung von globalen Ratingskalen unerläßlich (vgl. 1.5). Ziel jedes Rater-Trainings ist es, sicherzustellen, dass die Prüfer die Itemdeskriptoren verstehen, ähnlich interpretieren und vergleichbare Maßstäbe an die Performanz der Prüflinge anlegen. Das Spektrum möglicher

Trainingsmaßnahmen reicht von der Zusendung von Schulungsmaterialien in Form von Manualen und/oder Videomaterial an die Prüfer über E-Learning-Module bis zur Einladung der Prüfer zu Fortbildungen. Bei den beiden erstgenannten Vorgehensweisen ist es jedoch schwierig, den Erfolg der Trainingsmaßnahme sicherzustellen. Am zuverlässigsten läßt sich die Inter-Rater-Übereinstimmung mittels eines „Frame-of-reference-Trainings“ erhöhen. Dabei werden die Rater kalibriert in Bezug auf die Definitionen von Verhaltensdimensionen und aufgefordert, ihre Beurteilungen von Interaktionsbeispielen in einer moderierten Gruppendiskussion auszutauschen. In dem Training erhalten die Teilnehmer gezielte Rückmeldung über ihr Ratingverhalten, zudem werden Kriterien für gute und schlechte Leistungen in jeder Dimension etabliert. Der Gebrauch des Instruments sowie die Beobachtungsfertigkeiten werden so aktiv geübt. Eine weitere effektive Maßnahme zur Erhöhung der Inter-Rater-Reliabilität ist der Ausschluss von extrem abweichenden Prüfern (82, 83).

## 2.3 Entwicklung eines Rater-Trainings

Um die in 2.2 beschriebenen Ziele zu erreichen, wurde im Rahmen dieser Studie ein zweistündiges Rater-Training mit zwei Videobeispielen entwickelt, die eigens für diesen Zweck produziert worden sind: Ein Film zeigt eine Studienanfängerin („Negativ-Beispiel“), der zweite zeigt ein patientenzentriertes Gespräch auf Absolventenniveau („Positiv-Beispiel“). In beiden Videobeispielen handelt es sich um eine Simulationspatientin mit Bronchialasthma, die beim „Hausarzt“ vorstellig wird. Das negative Beispiel wurde mit einer Abiturientin gedreht, im Positivbeispiel übernahm ein ärztlicher Kollege und Interaktionsdozent die Rolle des Hausarztes. Beiden Darstellern wurde die Situation geschildert sowie einige Verhaltensanweisungen zur Verdeutlichung gegeben. Die Darstellung erfolgte dann spontan im Spiel mit der Simulationspatientin; es gab kein wörtliches Skript. Die Länge der Videos war an der Prüfungssituation orientiert und betrug maximal zehn Minuten. Aus zwei bis drei Filmaufnahmen wurde das jeweils beste Video ausgewählt für das Rater-Training. Die beiden Sequenzen wurden von den vier Forscherinnen (A. Fröhmel, I. Mühlinghaus, H. Ortwein & S. Scheffer) zunächst unabhängig voneinander ausgewertet im Hinblick auf positiv und negativ zu bewertendes Kommunikationsverhalten vor dem Hintergrund patientenorientierter Gesprächsführung, wie sie im Reformstudiengang Medizin vermittelt wird. Anschließend wurden die Kriterien zusammengetragen und diskutiert, um die Standards für das Rater-Training zu etablieren.

## 2.4 Durchführung des Rater-Trainings

Alle Personen, die im Rahmen der Studie Prüflinge mit der globalen Ratingskala beurteilen sollten, wurden schriftlich zum Rater-Training eingeladen und um Unterstützung des Forschungsvorhabens gebeten. Das Rater-Training hatten alle bis auf vier Prüfer besucht, die aus nicht zu ändernden Gründen nicht teilnehmen konnten. Die Teilnehmerzahl war auf zwölf Personen pro Training begrenzt; insgesamt wurden sechs Trainings in einem Zeitraum von zwei Wochen durchgeführt. Alle Beobachtergruppen - Prüfer, SPs und Experten - wurden separat trainiert, um die Situation bei einem späteren Einsatz des Instruments zu simulieren.

Jede Trainingssitzung wurde von jeweils zwei der vier an der Studie beteiligten Kolleginnen (A. Fröhmel, I. Mühlinghaus, H. Ortwein bzw. S. Scheffer) als Trainerinnen durchgeführt. Um Rollenkonflikte zu vermeiden, übernahm eine Person die Rolle des Moderators, während sich die andere inhaltlich an der Gruppendiskussion beteiligte. In der Schulung erhielten die Teilnehmer Informationen über das Instrument, das Prüfungsformat OSCE, die geplante Studie sowie das Ziel des Rater-Trainings. Um eine hohe Objektivität und Testfairness zu gewährleisten, wurden die Prüfer instruiert, bestimmte Verhaltensregeln einzuhalten (z. B. keine eigenen Fragen zur Aufgabe zu stellen und die Leistung des Prüflings nicht zu kommentieren). Verständnisfragen zu den Itembeschreibungen des Beurteilungsinstruments wurden geklärt. Anschließend wurden die zwei Videoaufnahmen von Arzt-Patienten-Gesprächen gezeigt, die als Benchmark dienten: Nach dem ersten Videobeispiel beurteilten die Teilnehmer unabhängig voneinander die kommunikativen Kompetenzen des Arztes auf dem Berliner Global Rating. Die Prüfer wurden instruiert, bei der Beurteilung darauf zu fokussieren, welche beobachtbare Wirkung das Verhalten des Arztes im Video auf den (Simulations-)Patienten hatte. In der sich anschließenden moderierten Gruppendiskussion wurden alle Teilnehmer gebeten, die Kriterien für ihre Einschätzung anhand beobachteter Beispiele zu erläutern. Falls ein Teilnehmer Vermutungen äußerte oder die Vergabe von Punkten mit „Sympathiewerten“ begründete, wurde vom Moderator nachgefragt, ob sich die Punktzahl durch im Video beobachtbares Verhalten untermauern lasse, ob beispielsweise durch beobachtbare Reaktionen des SPs erkennbar sei, wie das Verhalten des Arztes gewirkt habe. Damit wurde beabsichtigt, dass die Teilnehmer lernen, von Sympathiewerten und vermuteten Intentionen des schauspielernden Arztes zu abstrahieren. Die Beurteilungen wurden gesammelt und alle Kriterien zusammenfassend dargestellt. Von der Zielpunktzahl abweichende Rater wurden gefragt, ob sie sich in Anbetracht der von der Gruppe vorgetragenen Argumente dem Gruppenkonsens anschließen könnten. Gegebenenfalls wurde die Diskussion der Kriterien fortgeführt, um das Ziel der Standardisierung der Urteile durch Konsensfindung innerhalb der

Ratergruppe zu erreichen. Falls wichtige Aspekte gar nicht von den Teilnehmern genannt wurden, wurden diese von der zweiten Trainerin, die nicht moderierte, eingebracht. Danach wurde genauso mit dem zweiten Videobeispiel verfahren. Tabelle 2 zeigt die zwölf Schritte des Rater-Trainings.

**Tabelle 2: Ablauf des zweistündigen Rater-Trainings**

| | Inhalt | Dauer |
|---|---|---|
| 1. | Begrüßung und Vorstellung; Information über Ziel der Studie, des Rater-Trainings und Besonderheiten der Prüfungsmethode OSCE inkl. Verhaltensregeln für Prüfer | 20 min |
| 2. | Vorstellung des Berliner Global Rating, Besprechen der Itemdeskriptoren, Hinweise zum Ausfüllen, Klärung von Verständnisfragen | 10 min |
| 3. | Videobeispiel Nr. 1 (negativ) | 10 min |
| 4. | Individuelles Ausfüllen des Berliner Global Rating | 2,5 min |
| 5. | Nennen und Begründen der vergebenen Note durch jeden Teilnehmer, Beginn Konsensfindung | 15 min |
| 6. | Zusammenfassung der Argumente und evtl. Ergänzung durch Moderatoren | 5 min |
| 7. | ggf. Diskusssion und Abschluss Konsensfindung | 10 min |
| 8. | Videobeispiel Nr. 2 (positiv) | 10 min |
| 9. | Individuelles Ausfüllen des Berliner Global Rating | 2,5 min |
| 10. | Nennen und Begründen der vergebenen Note durch jeden Teilnehmer, Beginn Konsensfindung | 15 min |
| 11. | Zusammenfassung der Argumente und evtl. Ergänzung durch Moderatoren | 5 min |
| 12. | ggf. Diskusssion und Abschluss Konsensfindung; Verabschiedung | 15 min |
| | **Gesamtdauer Rater-Training** | **120 min** |

Falls ein Teilnehmer nicht bereit war, sich den mit der Gruppe erarbeiteten Beurteilungskriterien anzuschließen, wurde er von der Studie und damit der Prüfung ausgeschlossen. Dies war bei einem Simulationspatienten der Fall.

## 2.5 Stichprobe und Studiendesign

Die Studie wurde in den OSCE-Prüfungen des 3. und 5. Semesters des Reformstudiengangs Medizin an der Charité im Februar 2004 durchgeführt. Alle zur Prüfung zugelassenen

Studierenden wurden vor der Prüfung per Zufallsauswahl einer Prüfungsgruppe zugeteilt. 55 Studierende im 3. Semester und 58 Studierende im 5. Semester nahmen an der Prüfung teil; fünf Studierende fehlten am Tag der Prüfung. Die Stichprobe (n=113) bestand somit aus 42 Männern (37%) und 71 Frauen (63%) mit einem Altersdurchschnitt von 23 Jahren bei einer Spannweite von 20-40 Jahren. Die Verteilung von Alter und Geschlecht war in beiden Semestern vergleichbar (siehe Abbildung 2). Es gibt keine Hinweise darauf, dass diese beiden Kohorten sich hinsichtlich soziodemographischer Variablen von anderen Jahrgängen des Reformstudiengangs Medizin wesentlich unterscheiden. Es handelt sich bei der Stichprobe zwar um ein quasi-experimentelles Design, aber aufgrund der Größe der Stichprobe und der Einbeziehung von zwei vollständigen Kohorten spricht einiges dafür, dass die Stichprobe die Population annähernd repräsentativ abbildet.

**Tabelle 3: Stichprobe Studierende**

| | 3. Semester | 5. Semester | gesamt |
|---|---|---|---|
| N Stichprobe | 55 | 58 | 113 |
| N Frauen | 36 (65%) | 35 (60%) | 71 (63%) |
| Alter MW | 23,2 | 23,4 | 23,3 |
| Alter SD | 4,3 | 3,0 | 3,7 |

Die Studierenden dieser beiden Semester wurden darüber informiert, dass in einigen Stationen ein oder zwei weitere Personen anwesend sein würden und dass diese sowie die Prüfer und SP zu Studienzwecken ihre kommunikativen Kompetenzen beurteilen würden. Ihnen wurde versichert, dass diese Beurteilung nicht in die Note einfließen würde und dass die Beurteilung ihrer klinisch-praktischen Kompetenzen ausschließlich durch die Prüfer durchgeführt würde. Die Studierenden wurden instruiert, die zusätzlich anwesenden Personen zu ignorieren und sich wie gewohnt zu verhalten.

### 2.5.1 Studiendesign zu Hypothesen 1 und 2

Zur Beantwortung der Hypothesen 1 und 2 wurden die Prüflinge simultan von verschiedenen Beobachtergruppen auf dem (BGR) beurteilt. Das Studiendesign ist in Abbildung 2 graphisch dargestellt. In den Prüfungen wurden neben den klinisch-praktischen Fertigkeiten die kommunikativen Kompetenzen der Studierenden von jeweils drei verschiedenen Ratern - Prüfern, Experten und SPs - in Abbildung 2 symbolisiert durch die Ovale – auf dem Berliner

Global Rating bewertet. Insgesamt beurteilten 20 Prüfer, 16 Experten und 20 SPs die Prüflinge auf dem BGR.

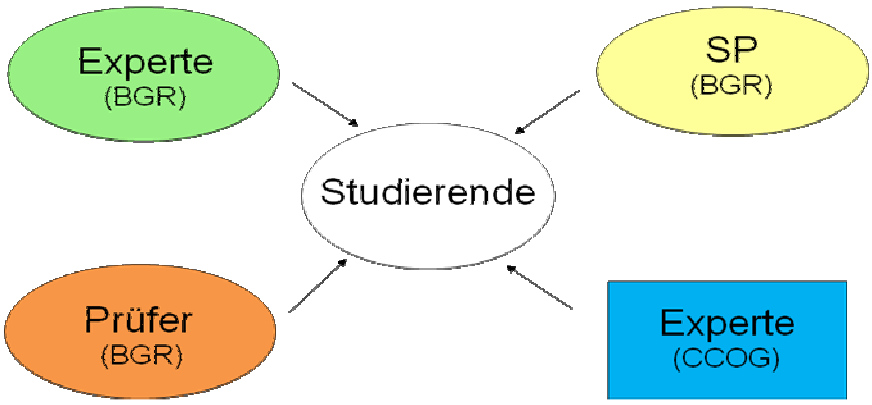


**Abbildung 2: Studiendesign.**

*Anmerkung:* BGR = Berliner Global Rating. CCOG = Checkliste

Dies geschah in Stationen mit Simulationspatienten-Kontakten, was auf drei von neun Stationen im 3. Semester und fünf von neun Stationen im 5. Semester zutraf. Die Aufgabenstellungen waren unabhängig von dieser Studie von Kollegen der Fächer, die in an den jeweiligen Themenblöcken des Semesters beteiligt sind, gestaltet worden, um die Lernziele des Blockes praktisch zu prüfen. Die in die Studie einbezogenen Stationen sind in Tabelle 4 dargestellt. In beiden Prüfungen hatten die Studierenden für jede Aufgabe acht Minuten zur Verfügung und eine Minute Zeit zum Wechseln. Für jeden Prüfling dauerte die Prüfung ca. 80 Minuten. Die Studierenden wurden in mehreren, nacheinander einbestellten Prüfungsgruppen à 8-10 Prüflingen parallel auf zwei Ebenen geprüft. Die Gesamtdauer der OSCE-Prüfung über alle Prüfungsgruppen betrug in Abhängigkeit der Anzahl der Prüflinge 5-6 Stunden.

**Tabelle 4: OSCE-Stationen der Studie**

| **3. Semester (N Prüflinge = 55)** | | |
|---|---|---|
| *Nr.* | *Aufgabe* | *Block* |
| 2 | Anamnese Entzündung | Entzündung/Abwehr |
| 4 | Sexualnamnese | Sexualität/Geschlechtsorgane/Hormone |
| 6 | Anamnese und Kniegelenk-untersuchung bei Rheuma | Entzündung/Abwehr |
| **5. Semester (N Prüflinge = 58)** | | |

| *Nr.* | *Aufgabe* | *Block* |
|---|---|---|
| 1 | Dermatologische Anamnese | Haut |
| 6 | Anamnese und Untersuchung der Augenbeweglichkeit | Sinnessysteme |
| 7 | Anamnese akute Hörminderung und Durchführung Rinne-Weber-Test | Sinnessysteme |
| 8 | Anamnese Streßsyndrom | Psyche: Erleben & Verhalten |
| 9 | Anamnese Panikattacke | Psyche: Erleben & Verhalten |

Die OSCE-Prüfer wurden von den für die Prüfung Verantwortlichen aus der Fakultät rekrutiert; sie waren Fachvertreter der in dem jeweiligen Semester geprüften Inhalte. Die Prüfer hatten wie in jeder OSCE-Prüfung die Aufgabe, die klinisch-praktischen Fertigkeiten der Prüflinge auf einer dreistufigen, stationsspezifischen Checkliste zu beurteilen. Zwei Prüfer der Stationen 8 bzw. 9 im 5. Semester rotierten im Verlauf der Prüfung zwischen den beiden Stationen. Mit Ausnahme von vier Prüfern hatten die Prüfer noch nie die Übung Interaktion unterrichtet. Im Rahmen der Studie gab es jedoch keine Möglichkeit, auf die Auswahl der Prüfer Einfluss zu nehmen und diese vier Prüfer durch andere Kollegen zu ersetzen. Den Prüfern sowie auch den SPs war bekannt, in welchem Semester sie prüften bzw. eingesetzt waren.

Als Expertengruppe mit hoher Kompetenz wurden erfahrene Interaktions-Dozenten definiert, die während mindestens zwei Semestern die Übung Interaktion unterrichtet hatten und die dafür notwendige Weiterqualifizierung durchlaufen hatten (vgl. 1.2.3). Dieser Personenkreis hat sich intensiv (z.T. auch in Weiterbildungen) mit dem Thema ärztliche Gesprächsführung beschäftigt und wurde eingeladen, an der Studie teilzunehmen. Da die erforderliche Zahl an Experten auf diesem Weg nicht zu erreichen war, wurde der Kreis auf Dozenten mit einem Semester Interaktions-Erfahrung, durchlaufener Weiterqualifizierung und guter Evaluation durch die Studierenden erweitert. Die Beurteilungen dieser Experten dienen als Kriterium („Gold-standard“). Kein Experte war Beurteiler in einem Semester eingesetzt, das er unterrichtet hatte, um Fehler in der Bewertung durch Konfundierung zu vermeiden. Den Experten wurde nicht mitgeteilt, welches Semester sie beurteilten.

Als Simulationspatienten wurden ausschließlich erfahrene Darsteller eingeteilt, die die jeweilige Rolle schon mehrfach dargestellt hatten sowie ein Grundlagen- und Aufbautraining zum Feedback-Geben absolviert hatten. Die Simulationspatienten erhielten den üblichen Stundensatz, die Teilnahme am Rater-Training wurde ebenfalls vergütet. Die Hälfte der SPs war mit Einsätzen in beide Prüfungen involviert.

### 2.5.2 Studiendesign zu Hypothese 3

Als weitere Validierungsstrategie sollen die Expertenurteile auf dem Berliner Global Rating mit Expertenurteilen auf dem in der Lehre eingesetzten Instrument (CCOG) verglichen werden (Hypothese 3, in Abbildung 2 symbolisiert durch das Rechteck). Daher beurteilten sechs weitere Experten die Studierenden mittels der CCOG-Checkliste. Für die Studie wurde eine Kurzfassung des CCOG in Form einer detaillierten Checkliste mit 28 Items und einer dreistufigen Bewertungsskala („ja", „teils", „nein") eingesetzt. Eine deutsche Fassung lag bereits vor (vgl. S. 85)3. Die Checkliste ist unterteilt in die Bereiche Gesprächsbeginn, Problemexploration, Verständnis für die Patientenperspektive, Struktur der Konsultation, Beziehungsaufbau und Konsultationsende. Bei zwei Items (Nr. 11: „klärte (unklare) Patientenäußerungen" und Nr. 22: „wenn er las, beeinträchtigte dies nicht den Dialog und die Beziehung zum Patienten") wurde als Antwortmöglichkeit „trifft nicht zu" ergänzt, da diese Items voraussichtlich in der Prüfungssituation nicht beobachtet werden können. Bei einem Item (Nr. 20: „hielt den OSCE-Zeitrahmen ein") wurde das Wort „OSCE" ergänzt, um Mißverständnisse zu vermeiden.

Die CCOG-Checkliste wurde von sechs weiteren Experten ausgefüllt, um Urteilsfehler durch das Bemühen um Konsistenz, das beim gleichzeitigen Ausfüllen von Globalem Rating und Checkliste durch denselben Experten auftreten könnte, zu vermeiden. Diese Beurteiler waren in allen Stationen des 3. Semesters und in drei zufällig ausgewählten Stationen des 5. Semesters (Nr. 1, 7 und 8) anwesend. Bei dieser Gruppe von Experten war besonders wichtig, dass sie mit dem CCOG sehr gut vertraut waren. Daher nahmen die Beurteilungen neben drei der an der Studie Beteiligten (A. Fröhmel, I. Mühlingahus und S. Scheffer) noch zwei sehr erfahrene SPs und eine langjährige Interaktionsdozentin vor.

### 2.5.3 Studiendesign zu Fragestellung 4

Zur Beurteilung der diskriminanten Validität (Fragestellung 4) sollen die in dieser OSCE-Prüfung gewonnenen Daten betreffend die klinisch-praktischen Fertigkeiten der Studierenden herangezogen werden, die die Prüfer auf der regulären OSCE - Checkliste beurteilen.

Jeder Rater füllte für jeden Studierenden ein Blatt mit dem Berliner Global Rating bzw. dem CCOG aus; bei den Prüfern war die Skala auf einem Bogen mit der Checkliste zu den klinisch-

[3] Übersetzung von Dr. med. Heiderose Ortwein mit Genehmigung der Autoren

praktischen Fertigkeiten abgedruckt. Alle Rater wurden instruiert, sich nicht über ihre Bewertungen der einzelnen Prüflinge auszutauschen. Jeder Studierende hatte einen Bogen mit selbstklebenden Etiketten, auf denen jeweils Vor- und Nachname, die Nummer der Prüfungsgruppe sowie die Startposition im Parcours vermerkt waren. Alle Rater hatten ebenfalls selbstklebende Etiketten mit einer dreistelligen Identifikationsnummer erhalten. Dadurch war jeder Bogen eindeutig zuzuordnen.

## 2.6 Statistische Analyse

Alle Bögen wurden zur weiteren Datenverarbeitung eingelesen mittels eines Beleglesegeräts und der Software FormPro, die die Korrektur von Lesefehlern erlaubt und die Daten als Text-Dateien ausgibt. Die Auswertung der Daten der Studie erfolgte mit SPSS 12.0 und SPSS 15.0 für Windows.

Insgesamt wurden im Rahmen der Studie 1.704 Beurteilungen auf einem der beiden Instrumente vorgenommen: von den verschiedenen Ratergruppen wurden 495 (3. Sem.) bzw. 870 (5. Sem.) Situationen mit dem Berliner Global Rating beurteilt, weitere 165 bzw. 174 Situations-Beurteilungen (3. bzw. 5. Sem.) wurden mittels der Kurzversion des Calgary-Cambridge-Observation Guide (CCOG) vorgenommen.

Die Analyse der Daten des Berliner Global Ratings zeigte, dass keine Beurteilergruppe fehlende Werte aufwies. Die Datenanalyse der CCOG-Checkliste ergab, dass zwei der 28 Items wie vermutet von den Ratern sehr häufig als nicht zutreffend eingeschätzt worden waren; die Rate der fehlenden Werte lag bei diesen Items bei bis zu 96%. Daher wurden Item 11 und Item 22 aus der Checkliste entfernt, so dass 26 Items verblieben. Desweiteren wurden Fälle mit mehr als drei fehlenden Datenpunkten aus der Analyse ausgeschlossen, um zu starke Verzerrungen zu vermeiden.

Voraussetzung für die Verwendung parametrischer Verfahren ist die Normalverteilung der Daten und das Vorliegen von Intervallskalenniveau. Zur Überprüfung der Verteilung der Daten wurden Histogramme und Q-Q-Plots herangezogen. Die optische Analyse der Datenverteilung mittels dieser beiden Verfahren unterstützt die Annahme der Normalverteilung bei den mit dem Berliner Global Rating und bei den mit der Checkliste erhobenen Daten. Um Fehlinterpretation zu vermeiden, wurden zusätzlich Schiefe und Kurtosis analysiert. Alle Schiefe- und Kurtosiswerte liegen im Bereich zwischen -1 und 1; lediglich in Station 6 weichen die Daten der Interaktions-

experten auf dem Berliner Global Rating leicht ab (Kurtosis = -1,3). Es kann daher von normalverteilten Daten ausgegangen werden.

Rating-Skalen erfüllen oft nicht die Voraussetzungen einer Intervallskala. Bei mehrstufigen Ratingskalen kann jedoch Intervallskalierung angenommen werden, wenn nur die Endpunkte der Skala beschrieben sind und diese Benennungen die bipolaren Extreme eines Kontinuums andeuten (81). Dies ist beim Berliner Global Rating der Fall. Bei der dreistufigen Checkliste kann kritisch diskutiert werden, ob von Intervallskalenniveau ausgegangen werden kann. In der Literatur wird kontrovers diskutiert, ob bei Rating-Skalen Verletzungen der Intervallskaleneigenschaften die Anwendung parametrischer Verfahren verbieten. Bortz & Döring (64) plädieren für einen pragmatischen Standpunkt, solange inhaltlich sinnvolle Ergebnisse aus der Anwendung resultieren. Nach Wirtz & Caspar (81) können Verfahren für intervallskalierte Daten auch dann zur Analyse von Ratingdaten angewendet werden, wenn die Äquidistanz der Skalenpunkte moderat verletzt ist. Im Zweifel empfehlen sie als zusätzliche Maßnahme, neben den Maßen für intervallskalierte Daten zusätzlich geeignete Maße für ordinalskalierte Daten anzugeben. Führen beide Maße zu ähnlichen Schlußfolgerungen, ist die Frage des Skalenniveaus nicht kritisch für die Interpretation. Dieser Argumentation möchte ich mich anschließen und verwende daher intervallskalierte Verfahren, deren Ergebnisse ich durch die zusätzliche Berechnung ordinalskalierter Maße absichere.

Mit den auf dem Berliner Global Rating erhobenen Daten wurde wie folgt verfahren: Für jeden Prüfling wurden für jede Station Gesamtwerte als Mittelwerte der vier Itemwerte des globalen Ratings pro Beobachter errechnet. Aus diesen Stationsergebnissen wurde ein Gesamtprüfungswert berechnet (Mittelwert der Stationswerte). Zur Beurteilung der Übereinstimmung der Prüfer- und SP-Urteile mit den Experten-Urteilen (Fragestellungen 1 & 2) ist die Berechnung von Produkt-Moment-Korrelationen nicht ausreichend, da diese nur die Ähnlichkeit der Werte im Sinne eines linearen Zusammenhangs berücksichtigt. Um einen bestehensrelevanten Einsatz zu rechtfertigen, müssen strengere Kriterien an die Übereinstimmung und eventuelle Niveauunterschiede angelegt werden. Daher werden zur Bestimmung der Kriteriumsvalidität (konkurrente Validität) exakte Übereinstimmungen in Form von Intraklassenkorrelationen (ICC) berechnet. Bedingung hierfür ist, dass für dasselbe Merkmal mehrere Meßwertreihen mit derselben Metrik vorliegen. Dies ist in dieser Studie der Fall. Da eine möglichst exakte Übereinstimmung interessiert, werden unjustierte ICC berechnet, so dass Mittelwertunterschiede zwischen Ratern zu Lasten der Güte des Zusammenhangs abgebildet werden (81). In der Literatur wird z. T. der Berechnungsmodus im Statistikprogramm SPSS kritisiert und der Intraklassenkorrelations-

koeffizient nach Lin (84) empfohlen, der einen Korrekturterm enthält. Zur Überprüfung wird daher auch der Koeffizient nach Lin berechnet. Bei geringer Differenz der Werte (< 0,15) wird der mit SPSS berechnete Koeffizient beibehalten. Zur weiteren Interpretation der Ergebnisse werden grafische Darstellungen nach Bland & Altman (85) erstellt. Diese Methode ist hilfreich, um die Übereinstimmung zwischen zwei Messreihen, in diesem Fall die Übereinstimmung mit dem Goldstandard, zu beurteilen.

Die Auswertung der CCOG-Checkliste erfolgte, indem für die Kategorie „ja“ 1 Punkt vergeben wurde, für die Kategorie „teils“ 0,5 Punkte und für die Kategorie „nein“ 0 Punkte. Für jeden Prüfling wurden pro Station die erreichten Punkte summiert, durch die Anzahl der Items dividiert und so ein Prozentwert des maximal möglichen Werts errechnet. Außerdem wurde ein Gesamtscore über alle Stationen berechnet als Mittelwert der Stationsergebnisse. In den Gesamtwert über alle Stationen gingen aufgrund des Ausschlusses von 24 Fällen wegen fehlender Daten 43 Fälle aus dem 3. Semester und 46 Datensätze aus dem 5. Semester ein. Zur Bestimmung der Konstruktvalidität (konvergente Validität, Hypothese 3) als Übereinstimmung der Expertenurteile auf zwei verschiedenen Beurteilungsinstrumenten werden Produkt-Moment-Korrelationen berechnet. Zusätzlich werden Spearman-Rho-Korrelationen als nicht-parametrische Maße bestimmt. Zur weiteren Interpretation der Ergebnisse wird der Determinationskoeffizient $r^2$ berechnet. Dieses Bestimmtheitsmaß definiert die Größe der Streuung von y, die durch x erklärt werden kann bzw. den Anteil gemeinsamer Varianz (64).

Zur Analyse der diskriminanten Validität (Fragestellung 4) werden die Prüfungsergebnisse der Studierenden im OSCE herangezogen. Die Prüfungsergebnisse wurden in Excel (Microsoft Office 2000) berechnet. Die Prüfer bewerteten in jeder Station auf einer stationsspezifischen Checkliste, wie gut die Studierenden die jeweilige Aufgabe erfüllten. Die Checklisten enthielten Items, die die für die Bewältigung der Aufgabenstellung relevanten Schritte stationsspezifisch beschreiben. Die Anzahl der Items variierte je nach Aufgabenstellung. Für jedes Item konnte die Bewertung „vollständig erfüllt“ (entspricht 1 Punkt), „teilweise erfüllt (entspricht 0,5 Punkten) oder „nicht erfüllt“ (entspricht 0 Punkten) vergeben werden. Für jeden Prüfling wurde von den Prüfern pro Station eine Checkliste ausgefüllt. Nach dem Einscannen der Daten wurde für jede Station die Gesamtpunktzahl errechnet und ein Prozentwert gebildet. Das individuelle Prüfungsergebnis ergibt sich aus dem Mittelwert der Stationsergebnisse. Darüber hinaus wurden Mittelwerte und Standardabweichungen über alle Prüflinge pro Station berechnet, um die Schwierigkeit der Stationen beurteilen zu können.

Die Normalverteilung der Prüfungsdaten wurde optisch mittels Q-Q-Plots und der Berechnung von Schiefe und Kurtosis überprüft. Alle Werte liegen im Bereich zwischen -1 und 1, so dass von normalverteilten Daten ausgegangen werden kann. Zur Frage der Intervallskaliertheit der Daten gelten dieselben Argumente wie bei der ebenfalls dreistufigen CCOG-Checkliste. Als Aspekt der diskriminanten Validität werden Produkt-Moment-Korrelationen zwischen den Ergebnissen der Studierenden auf dem Global Rating und ihren klinisch-praktischen Fertigkeiten (Prüfungsergebnis) berechnet sowie ebenfalls Determinationskoeffizienten. Zur Absicherung der Ergebnisse werden ebenfalls Spearman-Rho-Korrelationen bestimmt.

# 3 Ergebnisse

Im folgenden Teil wird zunächst ein Überblick über die Verteilung der Daten der Daten gegeben. Daran schließen sich die Ergebnisse der zur Beantwortung der Fragestellungen durchgeführten Analysen an.

## 3.1 Deskriptive Ergebnisse

### 3.1.1 Berliner Global Rating

Bezogen auf die gesamte Stichprobe wurden die Studierenden von der Expertengruppe im Mittel mit der Note 2,50 (SD = 0,52) bewertet, von der Prüfergruppe mit der Note 2,61 (SD = 0,56) und von der Gruppe der SPs mit der Note 2,37 (SD = 0,56).

Im 3. Semester betrugen die von den Experten vergebenen Werte im Mittel über alle Stationen die Note 2,61 (SD = 0,53). Die Prüfer waren im Mittel etwas strenger (2,75; SD = 0,64) und die SPs milder in ihrem Urteil (2,26; SD = 0,65). Im 5. Semester vergaben die Experten im Mittel über alle Stationen die Note 2,40 (SD = 0,50), während die Prüfer die Studierenden mit der Note 2,48 (SD = 0,44) und die SPs sie mit der Note 2,47 (SD = 0,45) beurteilten (vgl. Tabelle 5 und Tabelle 6). In Abbildung 1-3 sind die Verteilungen als Box-Whisker-Plots aufgetragen. Die detaillierten Einzelwerte der OSCE-Stationen mit Median, Perzentil 25 und 75, Minimum, Maximum, arithmetischem Mittel und Standardabweichung sind für jede Ratergruppe getrennt nach Semester in Tabelle 5 und Tabelle 6 dargestellt.

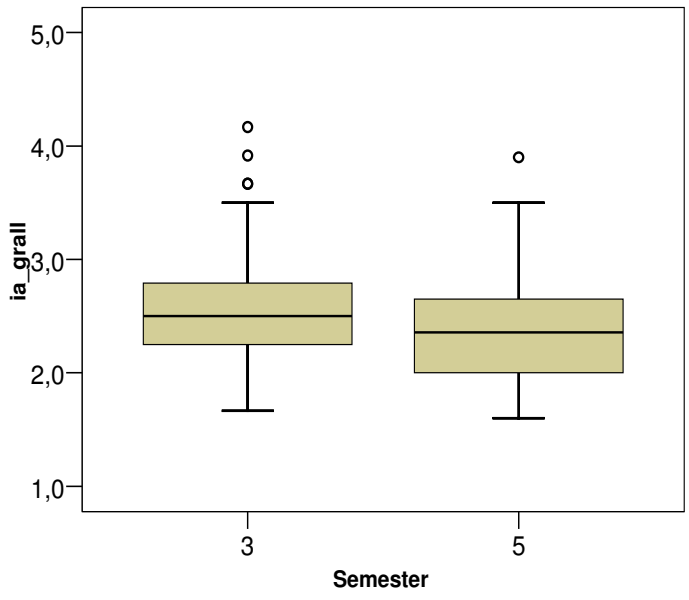

**Abbildung 3: Berliner Global Rating - Verteilung der Gesamtmittelwerte pro Semester in der Gruppe Experten.**

*Anmerkung:* Die Skala der Y-Achse reicht von Note 1-5, wobei 1 der bestmöglichen Leistung entspricht und 5 der schlechtesten.

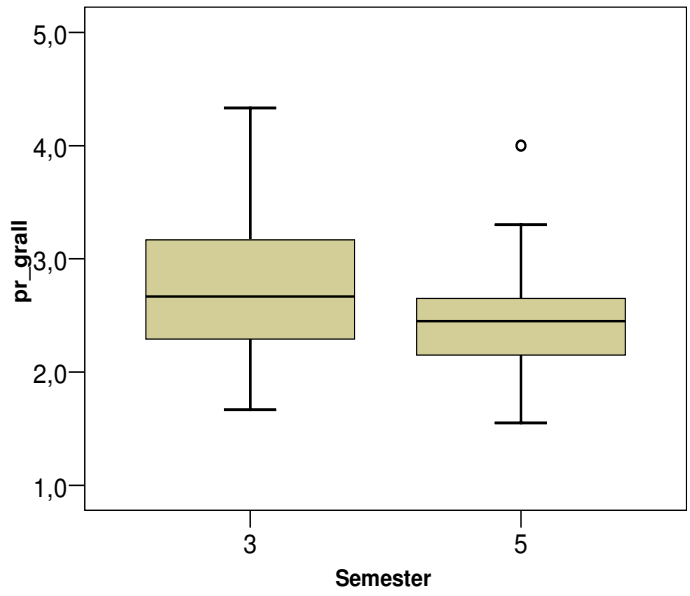


**Abbildung 4: Global Rating - Verteilung der Gesamtmittelwerte pro Semester in der Gruppe Prüfer**

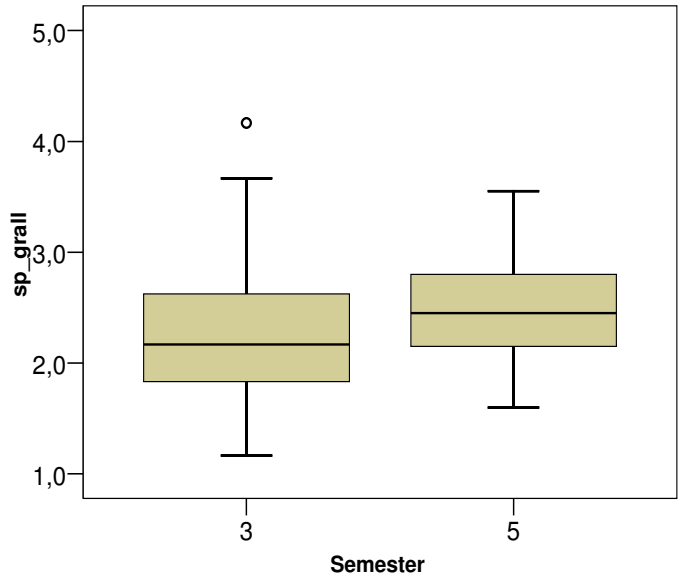


**Abbildung 5: Global Rating - Verteilung der Gesamtmittelwerte pro Semester in der Gruppe Gruppe SPs**

Wie die Box-and-Whisker-Plots zeigen, gibt es Unterschiede in der Verteilung der Daten zwischen den SPs und den anderen beiden Ratergruppen im 3. Semester; die SPs urteilten hier milder im Vergleich zu den Experten und Prüfern.

**Tabelle 5: Ergebnisse in den Einzelstationen nach Ratergruppen, 3. Semester**

| **3. Semester (n = 55)** | | | | | |
|---|---|---|---|---|---|
| | **Station Nr.** | **2** | **4** | **6** | **gesamt** |
| **Experten** | Median | 3,00 | 2,50 | 2,25 | 2,50 |
| | Percentil 25 | 2,50 | 2,25 | 1,50 | 2,25 |
| | Percentil 75 | 3,25 | 3,00 | 3,50 | 2,83 |
| | Min. | 1,00 | 1,00 | 1,00 | 1,67 |
| | Max. | 4,50 | 4,50 | 4,25 | 4,17 |
| | MW | 2,81 | 2,60 | 2,43 | 2,61 |
| | SD | 0,73 | 0,78 | 1,02 | 0,53 |
| **Prüfer** | Median | 3,00 | 2,50 | 2,75 | 2,67 |
| | Percentil 25 | 2,50 | 1,50 | 2,00 | 2,25 |
| | Percentil 75 | 3,75 | 3,25 | 3,50 | 3,25 |
| | Min. | 1,75 | 1,00 | 1,00 | 1,67 |
| | Max. | 4,50 | 4,75 | 5,00 | 4,33 |
| | MW | 3,11 | 2,41 | 2,74 | 2,75 |
| | SD | 0,71 | 1,03 | 0,93 | 0,64 |
| **SPs** | Median | 2,50 | 2,00 | 1,75 | 2,17 |
| | Percentil 25 | 2,00 | 1,50 | 1,25 | 1,83 |
| | Percentil 75 | 3,00 | 2,56 | 3,00 | 2,67 |
| | Min. | 1,25 | 1,00 | 1,00 | 1,17 |
| | Max. | 3,75 | 4,50 | 4,50 | 4,17 |
| | MW | 2,50 | 2,14 | 2,10 | 2,26 |
| | SD | 0,66 | 0,82 | 1,02 | 0,65 |

**Tabelle 6: Ergebnisse in den Einzelstationen nach Ratergruppen, 5. Semester**

| 5. Semester (n = 58) | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **Station Nr.** | **1** | **6** | **7** | **8** | **9** | **gesamt** |
| **Experten** | Median | 2,50 | 2,00 | 2,75 | 2,25 | 2,00 | 2,36 |
| | Percentil 25 | 2,00 | 1,50 | 2,25 | 1,69 | 1,50 | 2,00 |
| | Percentil 75 | 3,00 | 2,50 | 3,50 | 2,50 | 2,75 | 2,65 |
| | Min. | 1,25 | 1,00 | 1,00 | 1,00 | 1,00 | 1,60 |
| | Max. | 3,00 | 2,50 | 3,50 | 2,50 | 2,75 | 3,90 |
| | MW | 2,56 | 2,15 | 2,78 | 2,20 | 2,22 | 2,40 |
| | SD | 0,81 | 0,78 | 0,73 | 0,68 | 0,85 | 0,50 |
| **Prüfer** | Median | 2,25 | 2,75 | 2,75 | 2,38 | 2,00 | 2,45 |
| | Percentil 25 | 1,75 | 2,00 | 2,25 | 1,94 | 1,69 | 2,15 |
| | Percentil 75 | 2,75 | 3,25 | 3,06 | 3,06 | 2,56 | 2,69 |
| | Min. | 1,00 | 1,00 | 1,25 | 1,25 | 1,00 | 1,55 |
| | Max. | 3,75 | 4,50 | 4,00 | 4,50 | 4,75 | 4,00 |
| | MW | 2,31 | 2,67 | 2,63 | 2,52 | 2,24 | 2,48 |
| | SD | 0,66 | 0,81 | 0,59 | 0,81 | 0,83 | 0,44 |
| **SPs** | Median | 2,25 | 2,88 | 2,50 | 2,25 | 2,25 | 2,45 |
| | Percentil 25 | 2,00 | 2,25 | 2,00 | 1,75 | 1,75 | 2,15 |
| | Percentil 75 | 2,75 | 3,25 | 3,50 | 2,75 | 2,75 | 2,80 |
| | Min. | 1,00 | 1,25 | 1,00 | 1,00 | 1,00 | 1,60 |
| | Max. | 3,75 | 4,50 | 4,50 | 3,50 | 4,25 | 3,55 |
| | MW | 2,34 | 2,85 | 2,66 | 2,21 | 2,31 | 2,47 |
| | SD | 0,57 | 0,80 | 0,83 | 0,58 | 0,71 | 0,45 |

### 3.1.2 CCOG-Checkliste

Nach Ausschluß von Fällen mit mehr als drei fehlenden Werten verblieben 89 Fälle in der Stichprobe. Bezogen auf die gesamte Stichprobe betrug der Gesamtscore im Mittel 53,4% (SD = 12,2%) der maximal erreichbaren Punkte. Die Studierenden im 3. Semester (n = 43) erreichten im Mittel 48,7% der Punkte bei einer Standardabweichung von 10,5%. Der durchschnittliche Gesamtwert der Studierenden im 5. Semester (N = 46) lag im Mittel bei 57,7% (SD = 12,3%) bei einem Stichprobenumfang von 46 Prüflingen. Abbildung 6 zeigt die Verteilung der Daten. Die Ergebnisse in den einzelnen Stationen sind in Tabelle 7 detailliert dargestellt.

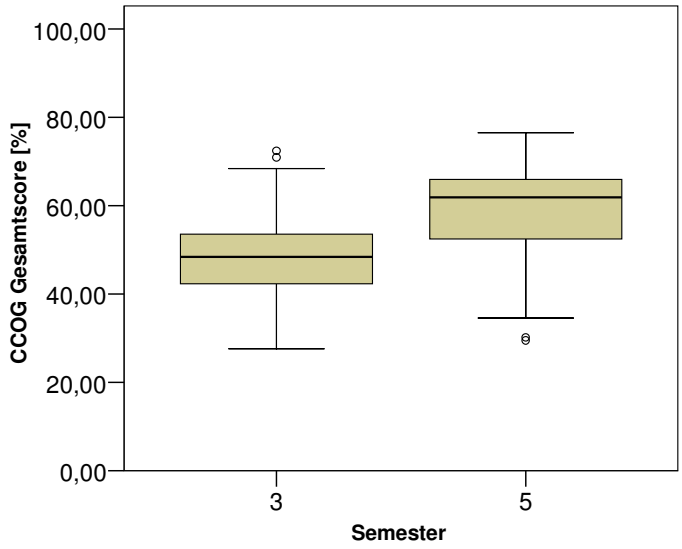


**Abbildung 6: CCOG - Verteilung der Gesamtwerte pro Semester**

**Tabelle 7: CCOG – Ergebnisse in den einzelnen Stationen**

| | 3. Semester | | | | 5. Semester | | | |
|---|---|---|---|---|---|---|---|---|
| **Station Nr.** | **2** | **4** | **6** | **gesamt** | **1** | **6** | **8** | **gesamt** |
| Median | 51,92 | 56,00 | 34,78 | 48,38 | 56,97 | 64,98 | 57,69 | 61,91 |
| Percentil 25 | 42,31 | 42,31 | 24,25 | 42,28 | 42,15 | 53,37 | 48,08 | 51,45 |
| Percentil 75 | 62,00 | 67,31 | 49,50 | 53,89 | 70,29 | 75,00 | 65,54 | 66,31 |
| Min. | 23,08 | 21,15 | 12,00 | 27,62 | 18,75 | 22,00 | 21,15 | 29,49 |
| Max. | 74,00 | 90,39 | 94,23 | 72,44 | 87,50 | 90,39 | 80,77 | 76,53 |
| MW | 51,13 | 56,07 | 39,28 | 48,72 | 56,08 | 62,10 | 55,62 | 57,74 |
| SD | 13,13 | 16,21 | 19,43 | 10,46 | 16,76 | 15,61 | 13,44 | 12,25 |
| N | 54 | 55 | 44 | 43 | 50 | 54 | 58 | 46 |

*Anmerkung:* Die Ergebnisse sind als Prozentwerte dargestellt.

## 3.2 Kriteriumsvalidität

Korrelationskoeffzienten können Werte von 0 bis 1 (bzw. -1) annehmen, wobei Null keinen Zusammenhang bedeutet und 1 bzw. -1 den größtmöglichen positiven bzw. negativen Zusammenhang. Bei der Quantifizierung der Validität durch Korrelationskoeffizienten sind Werte, die bedeutsam größer als Null sind und möglichst nahe bei 1 liegen, erstrebenswert. Zur Einordnung der Koeffizienten gelten Werte < 0.4 als gering, Werte zwischen 0.4 und 0.6 als mittelmäßig und Koeffizienten > 0.6 als hoch (64). Die berichteten Werte sind mit SPSS

berechnet. Der korrigierte Koeffizient nach Lin ergab lediglich marginale Differenzen im Bereich der zweiten Nachkommastelle, daher wird im weiteren auf die Darstellung verzichtet.

Die Übereinstimmung zwischen den Experten als Goldstandard und den OSCE-Prüfern ist gemäß Bortz & Döring (64) hoch, sie beträgt ICC = 0.74 bezogen auf die gesamte Stichprobe. Die Übereinstimmung zwischen Experten und SPs liegt mit ICC = 0.62 etwas niedriger, ist aber immer noch hoch. Tabelle 8 zeigt die Übereinstimmung zwischen den Experten und den beiden untersuchten Ratergruppen in der gesamten Stichprobe (n = 113) und in den einzelnen Semestern. Die exakte Übereinstimmung zwischen Experten und Prüfern ist in beiden Kohorten hoch (ICC = 0.70 und ICC = 0.78). Dies trifft auch auf Experten und SPs im 5. Semester zu mit ICC = 0.76, während die Übereinstimmung zwischen Experten und SPs im 3. Semester nur mittelmäßig ist (ICC = 0.54) und als einzige unter 0.7 liegt. Alle Korrelationen sind hoch signifikant ($p < 0.01$).

**Tabelle 8: Inter-Rater-Übereinstimmung zwischen Experten und Prüfern bzw. SPs**

| | | Experten x Prüfer | | Experten x SPs | |
|---|---|---|---|---|---|
| | **N** | **ICC** | **KI (95%)** | **ICC** | **KI (95%)** |
| **3. Semester** | 55 | 0.70** | 0.54-0.81 | 0.54** | 0.32-0.70 |
| **5. Semester** | 58 | 0.78** | 0.65-0.86 | 0.76** | 0.63-0.85 |
| **gesamt** | 113 | 0.74** | 0.65-0.82 | 0.62** | 0.49-0.72 |

Intraklassenkorrelationen (oneway). ** $p < .01$. KI = Konfidenzintervall.

In Tabelle 9 sind die Koeffizienten und Konfidenzintervalle (KI) für die Übereinstimmung zwischen Experten und OSCE-Prüfern detailliert für die einzelnen Stationen in jedem Semester dargestellt. Alle Korrelationen sind signifikant und liegen im mittleren bis hohen Bereich (0.37 - 0.69). Lediglich Station 2 im 3. Semester fällt aus diesem Gesamtbild heraus mit einer niedrigen Übereinstimmung (ICC = 0.25). Analog zeigt Tabelle 10 die Koeffizienten und Konfidenzintervalle für die Übereinstimmung zwischen Experten und SPs in den OSCE-Stationen. In fünf Stationen gab es mittelmäßige Übereinstimmungen der SPs mit den Experten (0.39 - 0.55), in zwei Stationen wurden nur niedrige Übereinstimmungen erreicht (0.25 in Station 2 und 0.27 in Station 8). In fünf der acht Stationen waren die Koeffizienten niedriger im Vergleich zu den Prüfern, in drei Stationen identisch bzw. annähernd gleich hoch. Die größten Differenzen zeigten

sich bei der Sexualanamnese (Nr. 4), der Anamnese bei Stresssyndrom (Nr. 8) und der Anamnese nach einer Panikattacke (Nr. 9).

**Tabelle 9: Inter-Rater-Übereinstimmung zwischen Experten und Prüfern pro Station**

| Nr. | Station | ICC | KI (95%) |
|---|---|---|---|
| 3. Semester | | | |
| 2 | Anamnese Entzündung | 0.25* | -0.01-0.48 |
| 4 | Sexualanamnese | 0.59** | 0.39-0.74 |
| 6 | Anamnese und Kniegelenks-untersuchung | 0.64** | 0.45-0.77 |
| 5. Semester | | | |
| 1 | Dermatologische Anamnese | 0.54** | 0.33-0.70 |
| 6 | Anamnese und Untersuchung der Augenbeweglichkeit | 0.51** | 0.29-0.71 |
| 7 | Anamnese akute Hörminderung und Durchführung Rinne-Weber-Test | 0.37** | 0.13-0.57 |
| 8 | Anamnese Streßsyndrom | 0.39** | 0.16-0.59 |
| 9 | Anamnese Panikattacke | 0.69** | 0.53-0.81 |

Intraklassenkorrelationen (oneway). * $p < 0.05$, ** $p < .01$. KI = Konfidenzintervall.

**Tabelle 10: Inter-Rater-Übereinstimmung zwischen Experten und SPs pro Station**

| Nr. | Station | ICC | KI (95%) |
|---|---|---|---|
| 3. Semester | | | |
| 2 | Anamnese Entzündung | 0.25* | -.01-.48 |
| 4 | Sexualanamnese | 0.45** | .21-.64 |
| 6 | Anamnese und Kniegelenksuntersuchung | 0.55** | .34-.71 |
| 5. Semester | | | |
| 1 | Dermatologische Anamnese | 0.46** | 0.23-0.64 |
| 6 | Anamnese und Untersuchung der Augenbeweglichkeit | 0.48** | 0.19-0.69 |
| 7 | Anamnese akute Hörminderung und Durchführung Rinne-Weber-Test | 0.39** | 0.15-0.59 |
| 8 | Anamnese Streßsyndrom | 0.27* | 0.01-.049 |
| 9 | Anamnese Panikattacke | 0.55** | 0.38-0.70 |

Intraklassenkorrelationen (oneway). * $p < 0.05$, ** $p < .01$. KI = Konfidenzintervall.

Für den Einsatz eines Prüfungsinstruments in bestehensrelevanten Prüfungen interessiert weiterhin, wie hoch die exakte Übereinstimmung zwischen Experten und Prüfern bzw. SPs in den Extremgruppen ist: Besteht Einigkeit zwischen Experten und Prüfern bzw. SPs, welche Leistungen besonders gut und welche eher schwach waren? Daher wurden zusätzlich Intraklassenkorrelationen für die 25% leistungsstärksten und -schwächsten Prüflinge berechnet. Die Quartile wurden anhand der Expertenurteile bestimmt.

Die Koeffizienten sind hoch, zum Teil sogar sehr hoch: Die Übereinstimmungen sind mit ICC = 0.86 (gesamte Stichprobe) sehr hoch zwischen Experten und Prüfern und hoch (ICC = 0.75) zwischen Experten und SPs. Während die Übereinstimmung zwischen Experten und Prüfern gleichmäßig hoch ist, zeigen sich Differenzen in der Übereinstimmung zwischen Experten und SPs in den verschiedenen Semestern (vgl. Tabelle 11).

**Tabelle 11: Übereinstimmung zwischen Ratergruppen bezogen auf das 1. und 4. Quartil**

| | | ICC | KI (95%) | N |
|---|---|---|---|---|
| **Experten x Prüfer** | 3. Semester | 0.83** | 0.67-0.92 | 26 |
| | 5. Semester | 0.85** | 0.70-0.92 | 29 |
| | gesamt | 0.86** | 0.77-0.91 | 55 |
| **Experten x SPs** | 3. Semester | 0.67** | 0.40-0.84 | 26 |
| | 5. Semester | 0.83** | 0.67-0.91 | 29 |
| | gesamt | 0.75** | 0.61-0.85 | 55 |

Intraklassenkorrelationen (oneway). ** $p < .01$. KI = Konfidenzintervall.

Tabelle 12 und Tabelle 13 zeigen Koeffizienten und Konfidenzintervalle für die Übereinstimmung zwischen Experten und Prüfern bzw. SPs in den Extremgruppen (1. und 4. Quartil). In der Gruppe der Prüfer sind fast alle Koeffizienten deutlich höher als in der gesamten Stichprobe und erreichen Werte im oberen Mittelbereich und darüber. Fünf der acht Stationen weisen im Extremgruppenbereich hohe Übereinstimmungen (> 0.6) auf.

In der Gruppe der SPs sind die Übereinstimmungen in vier Stationen höher in den Extremgruppen, allerdings sind zwei Koeffizienten im Vergleich zur gesamten Stichprobe nicht mehr signifikant. Auch liegt in dieser Gruppe kein Koeffizient > 0.7.

**Tabelle 12: Inter-Rater-Übereinstimmung zwischenExperten und Prüfern pro Station - Extremgruppen**

| Nr. | Station | ICC | KI (95%) |
|---|---|---|---|
| 3. Semester | | | |
| 2 | Anamnese Entzündung | 0.34* | -0.05-0.63 |
| 4 | Sexualanamnese | 0.65** | 0.37-0.83 |
| 6 | Anamnese und Kniegelenksuntersuchung | 0.79** | 0.59-0.90 |
| 5. Semester | | | |
| 1 | Dermatologische Anamnese | 0.72** | 0.50-0.86 |
| 6 | Anamnese und Untersuchung der Augenbeweglichkeit | 0.65** | 0.32-0.84 |
| 7 | Anamnese akute Hörminderung und Durchführung Rinne-Weber-Test | 0.50** | 0.18-0.73 |
| 8 | Anamnese Streßsyndrom | 0.33* | -0.03-0.62 |
| 9 | Anamnese Panikattacke | 0.69** | 0.45-0.84 |

Intraklassenkorrelationen (oneway). * $p < 0.05$, ** $p < .01$.

**Tabelle 13: Inter-Rater-Übereinstimmung zwischen Experten und SPs pro Station - Extremgruppen**

| Nr. | Station | ICC | KI (95%) |
|---|---|---|---|
| 3. Semester | | | |
| 2 | Anamnese Entzündung | 0.24 | -0.15-0.57 |
| 4 | Sexualanamnese | 0.58** | 0.26-0.79 |
| 6 | Anamnese und Kniegelenksuntersuchung | 0.68** | 0.41-0.84 |
| 5. Semester | | | |
| 1 | Dermatologische Anamnese | 0.45** | 0.11-0.70 |
| 6 | Anamnese und Untersuchung der Augenbeweglichkeit | 0.68** | 0.36-0.85 |
| 7 | Anamnese akute Hörminderung und Durchführung Rinne-Weber-Test | 0.67** | 0.41-0.83 |
| 8 | Anamnese Streßsyndrom | 0.18 | -0.19-0.51 |
| 9 | Anamnese Panikattacke | 0.62** | 0.34-0.80 |

Intraklassenkorrelationen (oneway). * $p < 0.05$, ** $p < .01$.

Zur weiteren Beurteilung der Übereinstimmung der Meßwerte wurden Bland & Altman-Plots erstellt. Dabei wurde auf der X-Achse der Mittelwert der Expertenurteile pro Prüfling als Goldstandard aufgetragen. Auf der Y-Achse wurde die Differenz der Mittelwerte (Y-X) von Prüfern und Experten bzw. SPs und Experten aufgetragen. In Abbildung 7 wurden die Urteile der Prüfer und Abbildung 8 die Urteile der SPs als Mittelwert Y zugrunde gelegt. Die horizontale Linie markiert die Nulllinie, d. h. bei allen Punkten, die auf dieser Linie liegen, ist die Beurteilung von Experten und Prüfern (bzw. Experten und SPs) identisch ausgefallen. Je weiter die Punkte von der Linie entfernt liegen, desto größer ist die Differenz der Prüfer- bzw. SP-Urteile zum Expertenurteil als Goldstandard.

Aufgrund der Skalierung des Berliner Global Rating bedeutet ein Wert im positiven Differenzbereich, dass die Prüfer bzw. SPs in diesem Fall strenger waren, während ein Wert im negativen Differenzbereich ein milderes Urteil im Vergleich zu den Experten anzeigt. Zur besseren Einordnung von Abweichungen sind jeweils zwei weitere Linien aufgetragen, um Abstände bis +/- 0,5 Notenstufen bzw. um +/- eine Note vom Expertenurteil zu kennzeichnen.

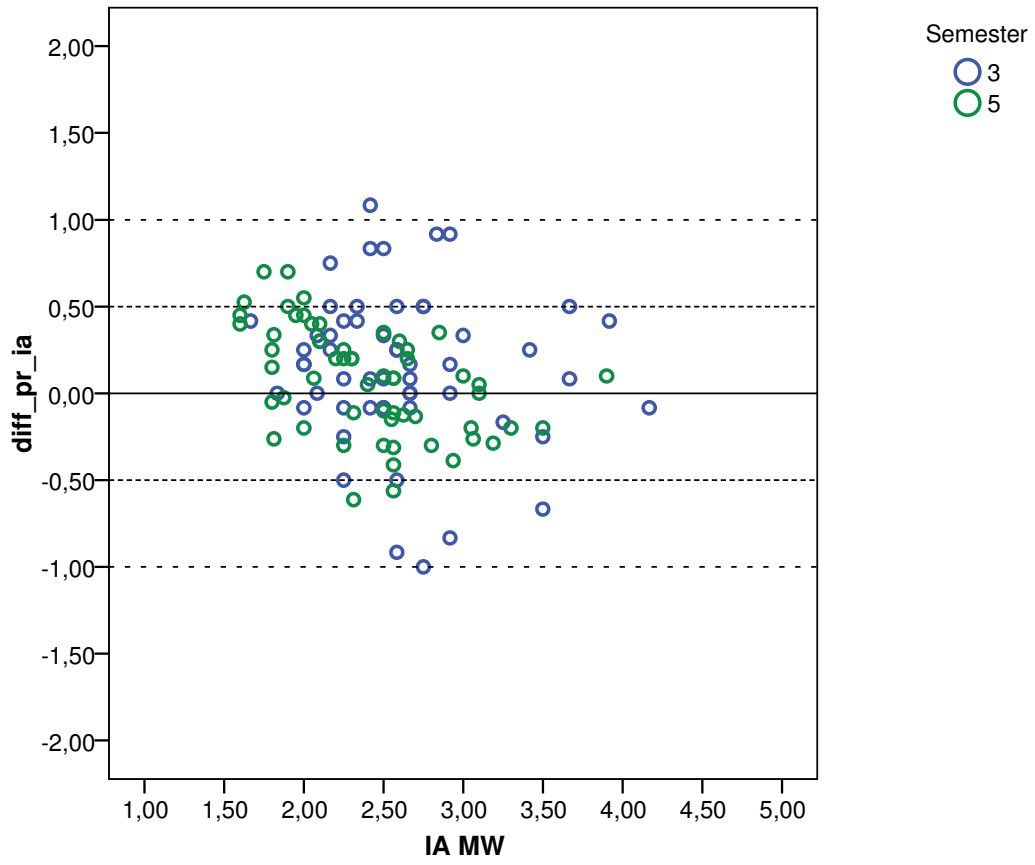


**Abbildung 7: Bland & Altman-Plot - Prüfer**

Wie die grafische Darstellung zeigt, streuen die meisten Werte im Bereich um maximal eine halbe Notenstufe um den Goldstandard; 85,9% der Prüfer-Urteile liegen in diesem Bereich. Mehr als die Hälfte der Prüfer-Urteile (52,7%) sind sogar nur höchstens eine Viertelnote vom Goldstandard entfernt. 5,3% der Prüfer-Urteile weichen mehr als eine halbe Note nach oben auf der Skala ab, in 8,8% der Fälle urteilten die Prüfer um mehr als eine halbe Note strenger als die Experten. Abweichungen um mehr als eine Note herum traten in weniger als 1% der Fälle auf.

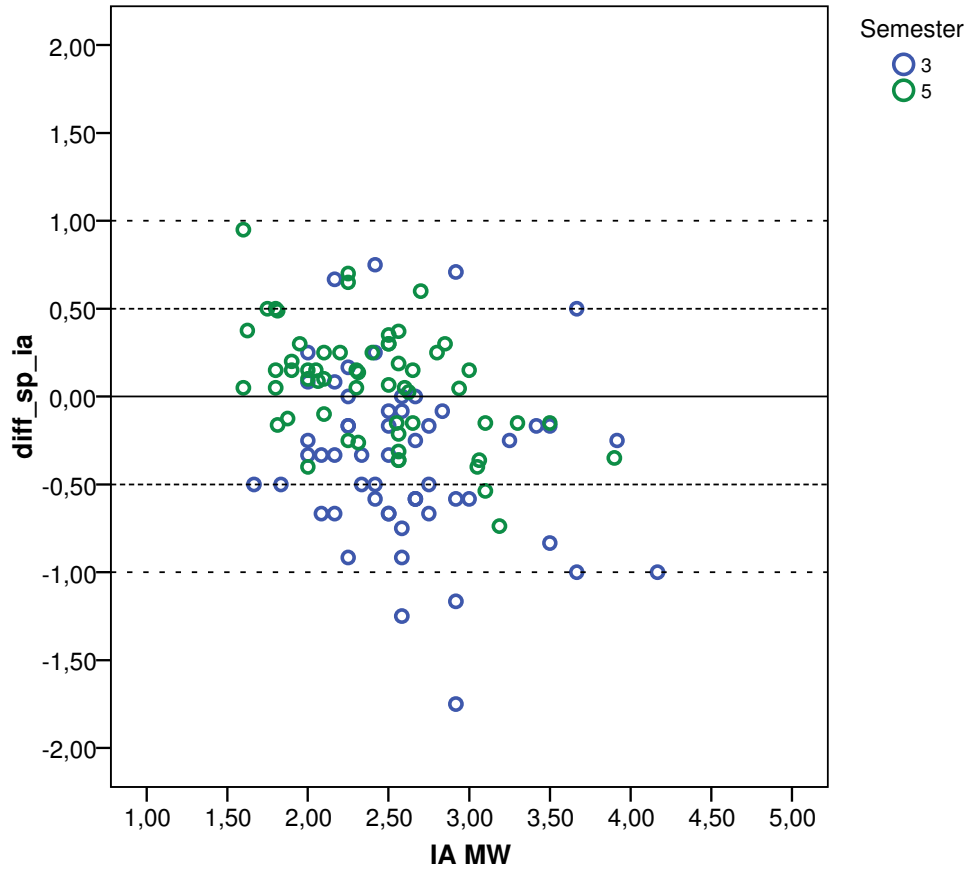


**Abbildung 8: Bland & Altman-Plot - SPs**

Anders gestaltet sich das Bild bei den SP-Urteilen. Wie schon der niedrigere Korrelationskoeffizient zeigte, ist die Übereinstimmung dieser Ratergruppe mit den Experten geringer: 74,3% der Urteile der SPs liegen im Bereich um maximal eine halbe Notenstufe höher oder niedriger im Vergleich zu den Expertenurteilen. Auffällig dabei ist die Konzentration im negativen Differenzbereich: 19,5% der Mittelwerte sind niedriger als die Experten-Mittelwerte, was bei der Skalierung einer als besser beurteilten Leistung entspricht. 6,2% der SP-Urteile sind

um bis zu einer halben Notenstufe strenger. Auch die Extremwerte sind zahlreicher und höher als in der Gruppe der Prüfer: 2,7% der SP-Mittelwerte weichen um mehr als eine Notenstufe von den Experten-Mittelwerten ab; in allen Fällen wurden die Studierenden als kompetenter beurteilt.

## 3.3 Konstruktvalidität

### 3.3.1 Konvergente Validität

Zur Bestimmung der konvergenten Validität als Zusammenhang der Meßergebnisse zwischen zwei verschiedenen Instrumenten wurden Produkt-Moment-Korrelationen berechnet. Die Korrelation zwischen den Expertenurteilen auf dem Berliner Global Rating und den Ergebnissen auf der CCOG-Checkliste ist hoch, sie beträgt r = -.72 bezogen auf die gesamte Stichprobe. Der negative Zusammenhang ist der entgegengesetzten Kodierung der Instrumente geschuldet. Tabelle 14 zeigt die Koeffizienten für die einzelnen Semester, beide Werte sind hoch, jedoch ist der Zusammenhang im 5. Semester deutlich stärker als im 3. Semester.

**Tabelle 14: Korrelationen zwischen Berliner Global Rating und CCOG-Checkliste**

| | **3. Semester** | **5. Semester** | **gesamt** |
|---|---|---|---|
| **r** | -0.61** | -0.82** | -0.72** |
| **n** | 43 | 46 | 89 |

** $p < 0.01$

Die zur Absicherung der Ergebnisinterpretation zusätzlich berechneten Rangkorrelationskoeffizienten (Spearmans Rho) sind etwas niedriger (rho = -.68 für die gesamte Stichprobe, rho = -.52 für das 3. Semester und rho = -.78 für das 5. Semester) als die Produkt-Moment-Korrelationen. Alle Werte sind wie die parametrischen Koeffizienten hochsignifikant und führen zu übereinstimmenden Schlußfolgerungen, so dass die Frage des Skalenniveaus nicht kritisch ist für die Interpretation und im Weiteren auf die parametrischen Koeffizienten Bezug genommen wird.

**Tabelle 15: Korrelationen zwischen Berliner Global Rating und CCOG-Checkliste**

| Nr. | Station | r | rho |
|---|---|---|---|
| 3. Semester | | | |
| 2 | Anamnese Entzündung | -0.56** | -0.54** |
| 4 | Sexualanamnese | -0.55** | -0.52** |
| 6 | Anamnese und Kniegelenksuntersuchung | -0.62** | -0.70** |
| 5. Semester | | | |
| 1 | Dermatologische Anamnese | -0.68** | -0.63** |
| 6 | Anamnese und Untersuchung der Augenbeweglichkeit | -0.35* | -0.39* |
| 8 | Anamnese Streßsyndrom | -0.57** | -0.56** |

* $p < 0.05$, ** $p < .01$.

Zur Interpretation der Ergebnisse wurde zudem der Determinationskoeffizient $r^2$ berechnet. Der Wert gibt an, in welchem Maße die Varianz einer Variablen durch die Varianz einer anderen Variablen bestimmt wird bzw. - da in dieser Studie kein Kausalmodell vorliegt - den Anteil der gemeinsamen Varianz zweier Merkmale an (64). $R^2$ variiert im Wertebereich zwischen 0 und 1, wobei 1 einen linearen Zusammenhang bedeutet. Werte $\geq 0.5$ gelten als hoch. In Abbildung 9 sind die Datenpaare als Streuungsdiagramm mit Regressionsgerade dargestellt. Der Verlauf der Geraden ergibt sich wieder aus der unterschiedlichen Kodierung der beiden Instrumente. $R^2$ beträgt 0.54, d. h. der Anteil der gemeinsamen Varianz der mit den beiden Meßinstrumenten erhobenen Werte beträgt 54%. Bei der Berechnung des Bestimmtheitsmaßes für die einzelnen Semester ergeben sich $r^2 = 0.37$ für das 3. Semester und $r^2 = 0.65$ für das 5. Semester.

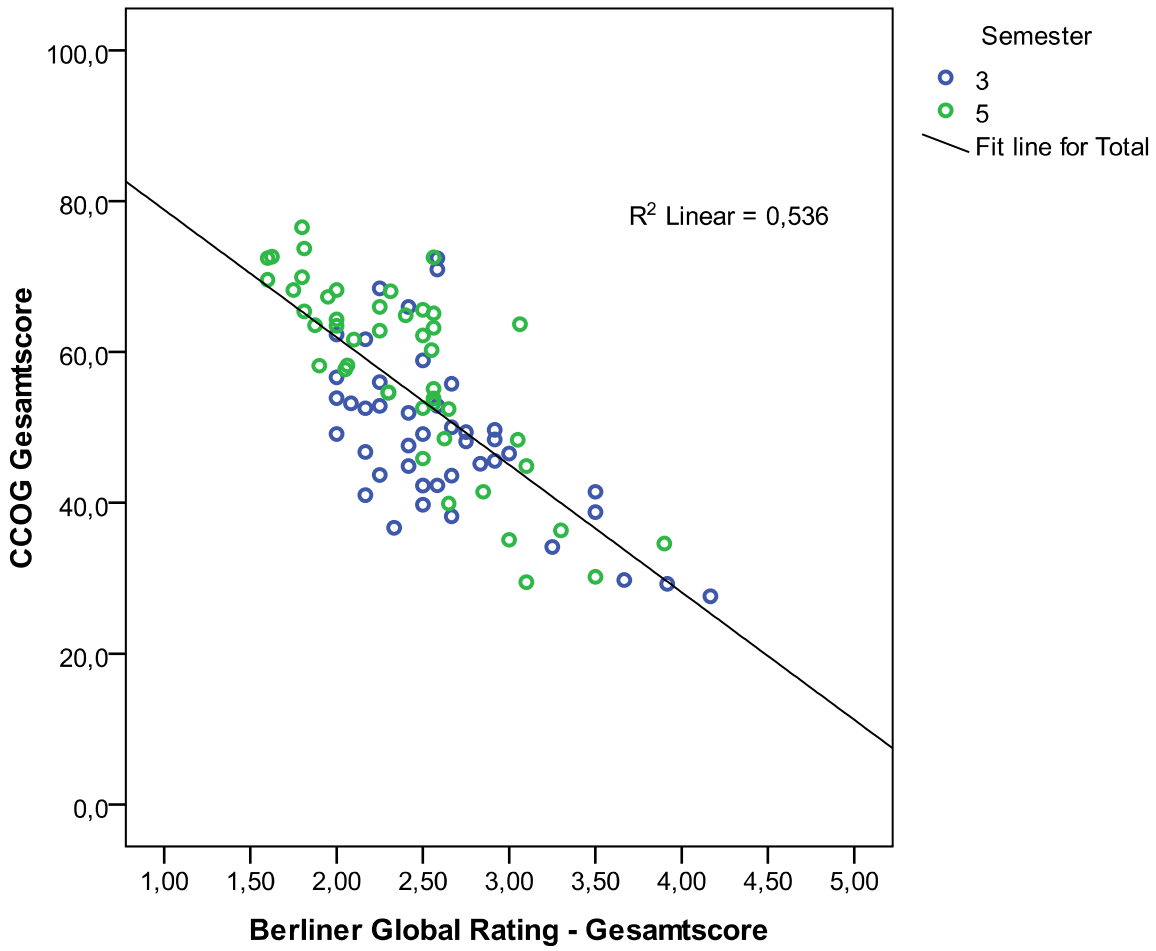


**Abbildung 9: Streuungsdiagramm Berliner Global Rating und CCOG**

### 3.3.2 Diskriminante Validität

In der OSCE-Prüfung beurteilten die Prüfer die klinisch-praktischen Fertigkeiten, die bei der Bewältigung der Aufgaben von den Studierenden demonstriert wurden, auf stationsspezifischen, dreistufigen Checklisten. Für jede Station wurde ein Prozentwert berechnet. Aus den Stationsergebnissen wurde der Mittelwert gebildet, der das Prüfungsergebnis darstellt. Studierende mit einem Ergebnis < 60% haben die Prüfung nicht bestanden. Dies traf auf zwei Prüflinge im 3. Semester und einen Prüfling im 5. Semester zu.

Die Leistungen der Studierenden in den OSCE-Prüfungen sind in Abbildung 10 und Abbildung 11 stationsweise dargestellt (Mittelwerte und Standardabweichungen). Aufgrund von Einsprüchen wurde auf Beschluss des Prüfungsausschusses in beiden Semestern jeweils eine Station

aus der Wertung genommen (im 3. Semester Nr. 9, Virologie, und im 5. Semester Nr. 5, Ophthalmoskopie).

Im 3. Semester waren die drei in die Studie einbezogenen Stationen am schwierigsten für die Prüflinge: In Station 2 (Anamnese bei Entzündung) erzielten die Studierenden im Mittel 51,4% der Punkte, gefolgt von Station 4 (Sexualanamnese; MW: 65,3%) und der Kniegelenksuntersuchung in Station 6 (MW: 86,6%).

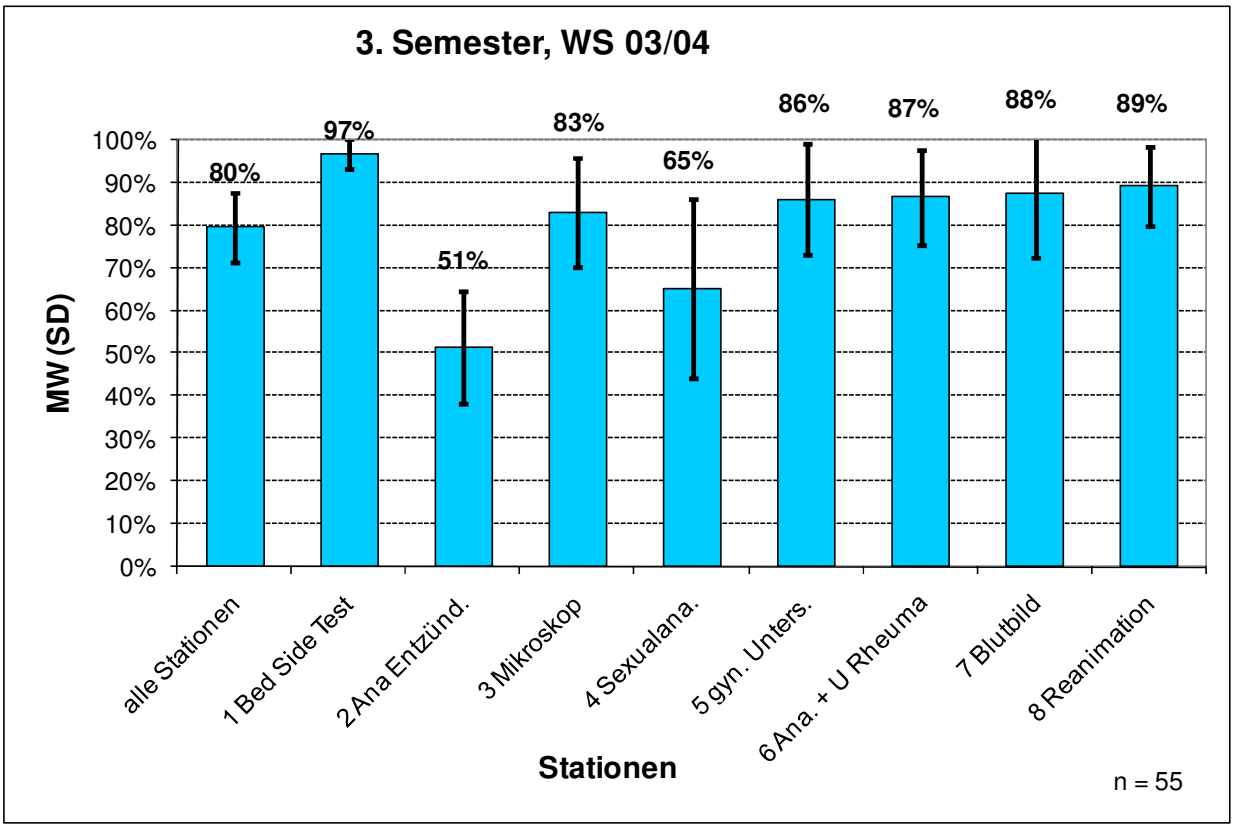


**Abbildung 10: Prüfungsergebnisse OSCE 3. Semester im WS 03/04**

Im 5. Semester erreichten die Studierenden bei den in die Studie eingegangenen Stationen im Mittel zwischen 56,6% (Nr. 8, Anamnese Streßsyndrom) und 80,9% (Nr. 7, Anamnese akute Hörminderung und Durchführung Rinne-Weber-Test) der Punkte.

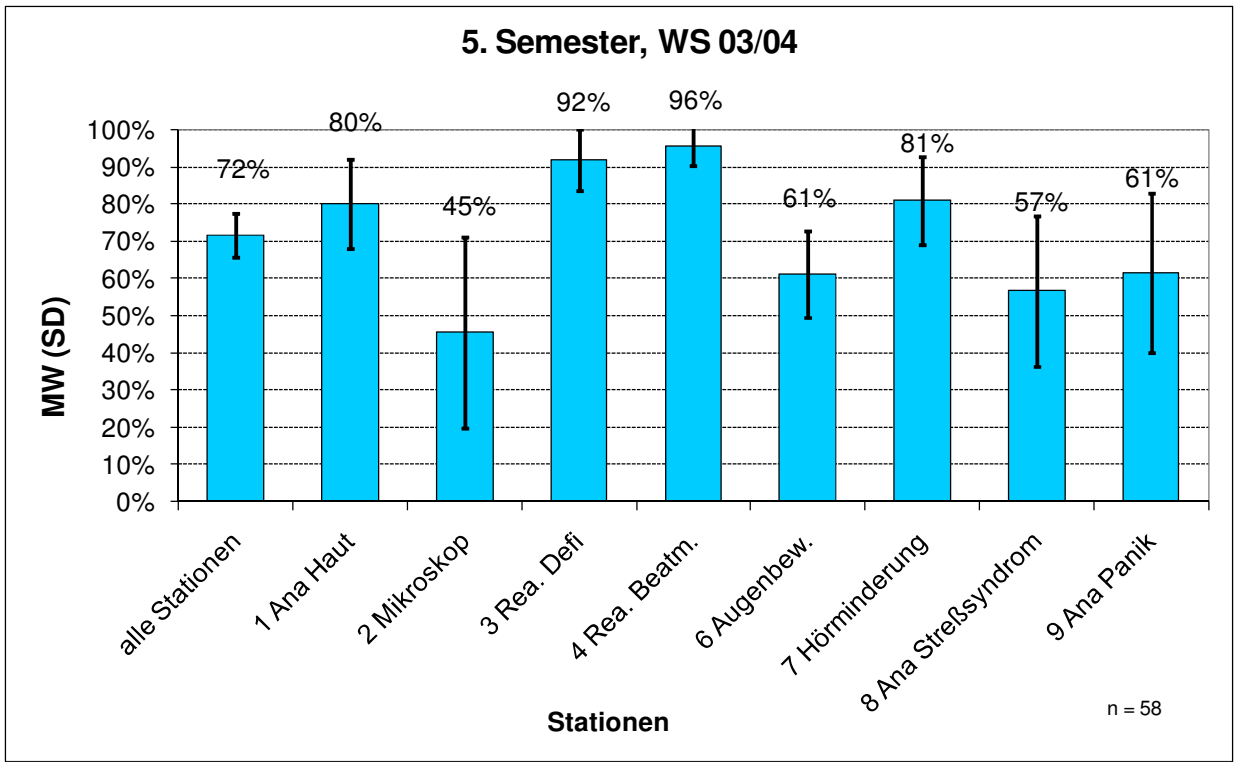


**Abbildung 11: Prüfungsergebnisse OSCE 5. Semester im WS 03/04**

Zur Untersuchung der diskrimanten Validität wurde der Zusammenhang zwischen dem Prüfungsergebnis in den einzelnen Stationen und der Beurteilung auf dem Berliner Global Rating durch die Experten bestimmt; es wurden Produkt-Moment-Korrelationen (r), Rangkorrelationen (rho) und Determinationskoeffizienten ($r^2$) berechnet.

**Tabelle 16: Berliner Global Rating und OSCE-Ergebnisse: Korrelationen und Determinationskoeffizient**

| Nr. | Station | r | rho | $r^2$ |
|---|---|---|---|---|
| 3. Semester | | | | |
| 2 | Anamnese Entzündung | -.17 | -.14 | 3% |
| 4 | Sexualanamnese | -.46** | -.42** | 22% |
| 6 | Anamnese und Kniegelenksuntersuchung | -.48** | -.50** | 23% |
| 5. Semester | | | | |
| 1 | Dermatologische Anamnese | -.40** | -.37** | 16% |
| 6 | Anamnese und Untersuchung der Augenbeweglichkeit | -.38* | -.32 | 15% |
| 7 | Anamnese akute Hörminderung und Durchführung Rinne-Weber-Test | -.05 | -.03 | 0% |
| 8 | Anamnese Streßsyndrom | -.33* | -.32* | 11% |
| 9 | Anamnese Panikattacke | -.62** | -.51** | 38% |

* $p < 0.05$, ** $p < .01$.

Die zur Absicherung der Ergebnisinterpretation zusätzlich berechneten Rangkorrelationskoeffizienten (Spearmans Rho) erbrachten mit den Produkt-Moment-Korrelationen übereinstimmende Ergebnisse, so dass die Verwendung der parametrischen Koeffizienten in der weiteren Ergebnisinterpretation unproblematisch ist.

Im Vergleich zu den Übereinstimmungen zwischen BGR und CCOG-Checkliste fallen die Zusammenhänge zwischen BGR und OSCE-Checkliste deutlich niedriger aus. In der Beziehung zwischen klinischen und kommunikativen Kompetenzen zeigen sich jedoch große Unterschiede zwischen den Stationen: In zwei Stationen gab es gar keinen Zusammenhang, in zwei weiteren einen schwachen Zusammenhang, in drei Stationen einen mittelmäßigen und in einer Station einen starken Zusammenhang (vgl. Tabelle 16; die negativen Vorzeichen ergeben sich aus der gegensätzlichen Kodierung). Der Determinationskoeffizient fällt jedoch durchweg niedrig aus und überspringt in keinem Fall die 50%-Grenze.

# 4 Diskussion

Ziel dieser Studie war es, ein geeignetes Instrument zur Prüfung kommunikativer Kompetenzen im Rahmen klinisch-praktischer Prüfungen (OSCE) von Medizinstudierenden zu identifizieren, ggf. zu adaptieren und zu überprüfen, inwieweit es unter Validitätsaspekten für einen Einsatz in Prüfungen geeignet ist. Ein globales Beurteilungsinstrument aus Kanada wurde im Rahmen dieser Studie übersetzt und adaptiert (80). Im Rahmen einer OSCE-Prüfung im 3. und 5. Semester des Reformstudiengangs Medizin beurteilten drei verschiedene, speziell geschulte Gruppen von Ratern (Experten, Prüfer, Simulationspatienten) die Studierenden hinsichtlich ihrer kommunikativen Kompetenzen auf dem Berliner Global Rating. Eine vierte Gruppe, ebenfalls aus Experten bestehend, beurteilte die Prüflinge auf einer Kurzfassung des Calgary-Cambridge Observation Guide, der CCOG-Checkliste.

## 4.1 Kriteriumsvalidität

Für die Prüfung der Arzt-Patienten-Interaktion existiert kein Goldstandard, gegen den die Validität eines Instruments getestet werden könnte. Daher wurde in dieser Arbeit als Validierungsstrategie die Beurteilung durch Experten angewendet. Der Vergleich der Prüfer- und Simulationspatientenurteile mit Expertenurteilen als Referenzkriterium erlaubt Aussagen über die konkurrente Validität.

### 4.1.1 Prüfer

Werden Prüflinge durch Fakultätsangehörige bewertet, findet sich oft ein „Deckeneffekt“, d. h. ein großer Teil der Bewertungen rangiert am oberen (positiven) Ende der Skala und die Daten enthalten wenig Varianz (86). Dadurch ist es schwierig, zwischen Studierenden mit guten und schwachen Leistungen zu unterscheiden. In der vorliegenden Studie ist kein Deckeneffekt nachweisbar: die Mediane bzw. Mittelwerte sind um den mittleren Wert der Skala verteilt ohne größere Schiefe, und das Spektrum der möglichen Noten wurde genutzt. Die Prüfer fällten im Mittel etwas strengere Urteile als die Experten (MW 2,61 vs. 2,50), was sich auch in den Mittelwerten pro Semester zeigt.

Die Daten zeigen hohe Übereinstimmungen zwischen Experten und Prüfern im Gesamtergebnis der Studierenden in beiden Semestern und erbringen damit gute Belege für die konkurrente Validität (Hypothese 1). In den Extremgruppen (1. und 4. Quartil) ist die Übereinstimmung mit

den Expertenurteilen höher als in der gesamten Stichprobe und mit Koeffizienten > 0.8 als sehr gut zu beurteilen. Dies zeigt, dass die Prüfer sehr gut in der Lage sind, exzellente und schwache Prüfungsleistungen zu identifizieren. Für den Einsatz des Instruments in bestehensrelevanten Prüfungen und die Sicherung eines minimalen Kompetenzlevels ist dieser Aspekt besonders wichtig.

Die Übereinstimmung des Gesamtmittelwerts pro Semester ist höher als die Koeffizienten in den einzelnen Stationen. Dies erklärt sich dadurch, dass sich bei der Berechnung eines Gesamtmittelwerts pro Prüfling die Beurteilungsfehler einzelner Rater ausmitteln. Die Zuverlässigkeit der Mittelwerte über mehrere Rater ist immer höher als die Zuverlässigkeit der Urteile eines einzelnen Raters (81).

Der Bland & Altmann-Plot verdeutlicht die guten Übereinstimmungen zwischen Experten- und Prüferurteilen in beiden Semestern: Über die Hälfte der Urteilspaare sind nicht mehr als eine Viertelnote voneinander entfernt; im Bereich um bis zu einer halben Note Differenz liegen 86% der Urteilspaare. Bei der Verwendung dieser grafischen Methode zur Interpretation gibt es in diesem Kontext keinen a priori definierten cut-off-point; Entscheidungen müssen inhaltlich begründet werden. Abweichungen über einer halben Notenstufe sind nicht ideal, aber im Vergleich zu anderen, wesentlich unzuverlässigeren Prüfungsmethoden gering und daher m. E. vertretbar. Eine Abweichung um mehr als eine Notenstufe ist jedoch erheblich und als problematisch einzustufen. Die Quote dieser Abweichungen lag jedoch unter 1% bei den Prüfern.

Den Inkongruenzen zwischen den Urteilen verschiedener Rater liegen v. a. zwei Ursachen zugrunde liegen: Sie können zum einen dadurch bedingt sein, dass die Rater nicht einig waren, welche Unteraspekte berücksichtigt werden müssen und wie diese in ein Gesamturteil zusammenzufügen sind. Bei komplexen Konstrukten wie dem in dieser Studie zu beurteilenden wird die dem Urteil zugrunde liegende operationale Definition jedes einzelnen Raters etwas unterschiedlich sein. Dies kann überprüft werden, indem die Rater befragt werden oder über ihre Urteile diskutieren. Implizit verwendete Zusatzregeln können so aufgedeckt werden. Die Konzeptionierung des Rater-Trainings zielt genau darauf ab. Indem die Rater ihre Urteile begründen müssen, werden Inkongruenzen sichtbar und diskutiert (vgl. 0).

Zum zweiten können Rater unterschiedliche Urteile fällen, weil sie die Ausprägung der Merkmale unterschiedlich einschätzen (81). Dem liegen häufig Beurteilungsfehler zugrunde, die zu

systematischen Verzerrungen führen. Es wird zwischen vier Arten von Urteilsfehlern unterschieden:

1. Wahrnehmungsfehler (z.B. Halo-Effekt, Primay-Recency-Effekte)
2. Deutungs- und Interpretationsfehler (z. B. Milde-/Strengefehler, Tendenz zur Mitte)
3. Erinnerungsfehler (z. B. Selektionsfehler)
4. Wiedergabefehler (z. B. durch Konformitätsdruck)

Der Halo-Effekt beschreibt die Tendenz, von einem beobachteten Merkmal fehlerhaft auf andere Merkmale zu schließen (z. B. kann so sprachliche Gewandtheit zu einem überhöhten Kompetenzeindruck führen). Primacy-Recency-Effekte bezeichnen Reihungseffekte, aus denen Kontrastfehler entstehen, z. B. wenn der erste Prüfling sehr leistungsstark oder –schwach war. Aufgrund der Struktur der Prüfung kann die Reihenfolge nicht beeinflußt werden. Primacy-Recency-Effekte können reduziert werden, indem den Prüfern erlaubt wird, die Bewertung der ersten ein bis zwei Prüflinge nachträglich zu verändern. Aus diesem Grund wurden die an der Studie beteiligten Rater wurden im Rater-Training explizit auf diese Möglichkeit hingewiesen.

Beim Milde- oder Strengefehler werden Personen systematisch zu positiv oder zu negativ beurteilt. Im ersten Fall würde dies zu Deckeneffekten führen. Strengefehler können auftreten, wenn Prüfer der Meinung sind, Studierenden „eine Lektion erteilen“ zu müssen. Das Rater-Training bietet die Chance, solche destruktiven Haltungen zu identifizieren und diese Überzeugungen infrage zu stellen. Im Extremfall wäre der Ausschluss des Prüfers bzw. die Herausnahme der Beurteilungen aus der Wertung angemessen, um eine faire Prüfung zu sichern.

Eine Häufung der Urteile im mittleren Bereich (Tendenz zur Mitte) tritt vor allem auf, wenn die Skalen nicht mit Extrembeispielen verankert worden sind. Dem wurde mit der Auswahl der Videobeispiele im Rater-Training entgegengewirkt. Dadurch lernten die Rater realistische Beispiele für die höchsten und niedrigsten Werte kennen.

Erinnerungsfehlern kann durch regelmäßige Pausen für die Prüfer bzw. Auswechseln der Prüfer vorgebeugt werden. Wiedergabefehler sind eher unwahrscheinlich, da die Studienteilnehmer instruiert waren, sich nicht über die Bewertungen auszutauschen. In der regulären OSCE-Prüfung wäre der Prüfer allein Urteilender, so dass kein Konformitätsdruck entstehen kann.

Da die wahren Merkmalsausprägungen unbekannt sind, ist es schwierig, Urteilsfehler zu identifizieren (64). Die meisten Urteilsfehler lassen sich durch Aufklärung über ihr Vorkommen und ihre Effekte verringern. Dafür bietet das Rater-Training eine hervorragende Gelegenheit;

darüberhinaus erlaubt es die Identifikation stark abweichender Rater. Diese müßten entweder intensiver trainiert werden oder im Extremfall ausgeschlossen werden. Generell fallen Urteilsfehler weniger ins Gewicht, wenn sich das Gesamturteil aus mehreren Einzelurteilen zusammensetzt. Dies zeigt sich in den höheren Übereinstimmungen über mehrere Stationen.

Absolut betrachtet erhielten die Studierenden im 5. Semester im Mittel von Experten und Prüfern bessere Noten. Die Interpretation dieser plausibel erscheinenden Werte ist jedoch schwierig, da in den beiden Semestern unterschiedliche Aufgabenstellungen beurteilt wurden und die Prüfer und Experten in beiden OSCE-Prüfungen nicht identisch waren. Um einen Leistungszuwachs über die Studiendauer nachzuweisen, der auch ein Beleg für die Konstruktvalidität des Berliner Global Rating wäre, ist ein anderes Studiendesign erforderlich: Den Teilnehmern unterschiedlicher Semester müßten identische Aufgaben gestellt werden, die von denselben Prüfern (bzw. Experten) ohne Kenntnis der Ausbildungsstufe beurteilt werden. Bessere Leistungen fortgeschrittener Studierender wären zu erwarten. Dies wurde in der Originalstudie von Hodges & McIlroy (80) für das Instrument untersucht und gezeigt. Im Rahmen der vorliegenden Arbeit war aufgrund der Fragestellungen ein Design notwendig, mit dem dieser Aspekt der Konstruktvalidität nicht untersucht werden konnte und mit dem die Unterschiede zwischen den Semestern nur sehr vorsichtig interpretiert werden können.

Aufgrund des Charakters der Studie als Felduntersuchung kam es zu Einschränkungen in der Umsetzbarkeit des Studiendesigns, deren Einfluß im Folgenden genauer analysiert werden soll. Eine Einschränkung bestand darin, dass vier Prüfer nicht wie vorgesehen am Rater-Training teilnehmen konnten. Dies betraf die Stationen Sexualanamnese (Nr. 4, 3. Semester), Anamnese Streßsyndrom und Anamnese Panikattacke (Nr. 8 bzw. 9, 5. Semester).

Die Analyse der Daten unter Ausschluss dieser untrainierten Prüfer zeigte im 5. Semester eine leicht höhere (Streßsyndrom, ICC = .44) bzw. minimal niedrigere Übereinstimmung zwischen den verbleibenden trainierten Prüfern und Experten (Panikattacke, ICC = .67). Im 3. Semester führte der Ausschluss des untrainierten Prüfers dagegen zu einer deutlich niedrigeren Übereinstimmung (Sexualanamnese, ICC =.39).

Für dieses unerwartete Ergebnis gibt es zwei Erklärungsansätze: Die vier untrainierten Prüfer waren zufällig Kollegen, die in den Fächern Psychiatrie, Psychosomatik bzw. Sexualmedizin tätig sind und durch Aus- oder Weiterbildungen vertiefte Kenntnisse der Gesprächsführung (im

psychotherapeutischen Setting) hatten. Es kann angenommen werden, dass diese Prüfer in der Interpretation der Items des Beurteilungsinstruments weniger Schwierigkeiten hatten als andere Prüfer. Drei der vier Prüfer (alle 5. Semester) hatten zudem schon einmal die Übung Interaktion unterrichtet. Der verbleibende Prüfer im 3. Semester war hingegen schon im Training durch Schwierigkeiten bei der Beurteilung der Trainingsvideos aufgefallen, so dass die Herausnahme des zweiten Prüfers, obwohl untrainiert, zu einer geringeren Übereinstimmung führte.

Im Rahmen der Studie gab es keine Möglichkeit, für die Prüfung eingeteilte Personen auszutauschen. Bei einem bestehensrelevanten Einsatz des Prüfungsinstruments sollten mehrere Strategien zum Einsatz kommen, wenn während des Trainings Schwierigkeiten im Umgang mit dem Beurteilungsinstrument beobachtbar sind: Gegebenenfalls sollte diese Person noch intensiver geschult werden. Ist dies nicht möglich, wäre es wichtig, auf Ersatzprüfer zurückgreifen zu können, oder in dieser Station sollte auf die Prüfung der kommunikativen Kompetenzen verzichtet werden.

Eine zweite Einschränkung bestand darin, dass vier Fakultätsangehörige als Prüfer im 5. Semester (Stationen 8 & 9) involviert waren, die zuvor schon als Interaktionsdozenten tätig waren. Drei davon konnten nicht am Rater-Training teilnehmen (s. o.). Dies könnte dazu geführt haben, dass die Ergebnisse künstlich erhöht worden sind. Um dies zu überprüfen, wurde die Übereinstimmung zwischen Experten und Prüfern über alle Stationen neu berechnet, unter Ausschluss der Stationen 8 & 9. Die Übereinstimmung ist ähnlich hoch (ICC = .75) wie der in 3.2 berichtete Wert (ICC = .78), in den alle fünf Stationen eingeflossen sind, so dass eine Verzerrung der Ergebnisse verneint werden kann.

### 4.1.2 Simulationspatienten

In dieser Arbeit wurde als weiterer Aspekt der Kriteriumsvalidität untersucht, wie gut die Urteile der Simulationspatienten mit denen der Experten übereinstimmen (Hypothese 2). Dabei zeigte sich insgesamt eine gute Übereinstimmung der SP-Bewertungen mit den Experten, die jedoch immer hinter derjenigen der Prüfer mit den Experten zurückblieb.

Dies lag v.a. an der mittelmäßigen Übereinstimmung der Raterurteile im 3. Semester, die nur in den Extrembereichen hoch war. Die Analyse zeigt, dass die Studierenden in Station 4 (Sexualanamnese) von den SPs deutlich milder beurteilt wurden als von den Experten (MW 2.14 vs. 2.60) bei gleichem Range (1.0 – 4.5). Dies traf etwas weniger ausgeprägt auch auf Station 6 (Anamnese und Kniegelenksuntersuchung) zu. Die höhere Übereinstimmung im 5. Semester

könnte zum Teil auf Trainingseffekte zurückzuführen sein: Die OSCE-Prüfung des 3. Semesters hatte am Vortag stattgefunden. Mehrere SPs waren auch dort eingesetzt und daher am zweiten Tag möglicherweise geübter im Umgang mit dem Global Rating. Außerdem legten sie im direkten Vergleich der Studierenden zu verschiedenen Ausbildungszeitpunkten möglicherweise strengere Maßstäbe an die Prüflinge im 5. Semester an.

Wie der Bland & Altman-Plot (Abbildung 8) deutlich zeigt, fielen die differierenden Urteile der SPs meistens zugunsten der Studierenden aus. Größere Abweichungen um mehr als eine halbe Notenstufe gab es nur im Sinne besserer Noten. Der Befund, dass die SP milder als die beiden anderen Gruppen urteilten, ist konsistent mit anderen Forschungsergebnissen: Meßfehler von SP fallen in der Regel zugunsten der Studierenden aus (87). Für die milderen Urteile der SPs gibt es zwei Erklärungsansätze, die möglicherweise kombiniert auftraten:

Ein Grund für die mildere Bewertung könnte sein, dass die SPs im Gegensatz zu den Prüfern (und Experten) über einen Zeitraum urteilen mußten, in dem sie selbst aktiv waren und sich auf die standardisierte Darstellung der Rolle konzentrieren mußten. Diese Ablenkung könnte ihre Erinnerung an Details der Gesprächsführung beeinträchtigt haben. Möglicherweise haben daher Primacy- und Recency-Effekte zu einem positiveren Urteil beigetragen: Dabei überstrahlen extreme Merkmalsausprägungen am Anfang oder Ende die nachfolgenden bzw. zuvor liegenden Ausprägungen und dominieren das Urteil (64). Die Studierenden des Reformstudiengangs üben während des Unterrichts in der Übung Interaktion verschiedene Gesprächsformen. Begrüßung, Vorstellung und Gesprächsbeginn sowie der Gesprächsabschluss und die Verabschiedung sind jedoch in jedem Gespräch wichtig und werden daher wiederholt trainiert. Man kann daher davon ausgehen, dass die Studierenden diese Aspekte relativ gut beherrschen; dies zeigt sich auch in der Erfolgsquote bei entsprechenden Items in OSCE-Checklisten. Die beschriebenen Urteilsfehler können daher zu einer positiveren Bewertung geführt haben.

Ein anderer Grund könnte im Rollenverständnis der SPs liegen: Die Darsteller sind vorrangig im Unterricht tätig, wo ein Schwerpunkt ihrer Tätigkeit neben der Darstellung einer Rolle das Geben von konstruktivem Feedback ist. Die Darsteller sehen ihre Hauptaufgabe darin, zum Lernerfolg der Studierenden beizutragen. Das Prüfen von Studierenden gehört sonst nicht zu ihren Aufgaben. Möglicherweise fiel es ihnen daher schwerer als den Prüfern, schlechte Noten zu vergeben. Anderen Autoren zufolge geben SPs im Zweifel den Studierenden eher eine bessere Beurteilung, weil sie die Karriere eines Prüflings nicht durch ihr Urteil gefährden wollen (87). Um die zu milden Urteile der SPs zu korrigieren, scheinen zwei Maßnahmen geeignet: Das Training könnte bei dieser Ratergruppe intensiviert werden. Im Training der SPs sollte verstärkt

auf den Mildefehler hingewiesen, die Befürchtungen hinsichtlich der Auswirkungen strenger Urteile thematisiert und insbesondere bei dem Videobeispiel mit der schwachen Leistung auf der Vergabe entsprechend schlechter Noten bestanden werden. Die Tatsache, dass die Übereinstimmung zwischen Experten und SPs im 5. Semester fast genauso hoch war wie die der Prüfer mit den Experten, was als Trainingseffekt bei mehrfach eingesetzten SPs gedeutet werden kann, spricht für eine Intensivierung des Rater-Trainings. Ein weitere Maßnahme besteht im Ausschluss von SPs, die extrem urteilen: Fallen im Training bei einem SP – analoges gilt natürlich auch für Prüfer – extreme Mildefehler (oder Strengefehler) auf, sollte diese Person nicht an der Prüfung beteiligt werden (63).

Zusammenfassend kann zur Beantwortung von Hypothese 1 und 2 Folgendes festgehalten werden: Es wird oft argumentiert, dass es Fakultätsangehörigen schwer fällt, Sachverhalte zu beurteilen, auf die im beruflichen Alltag wenig Wert gelegt wird[4] (88). Daraus wurde zum Teil abgeleitet, dass die Prüfung kommunikativer Kompetenzen durch Fakultätsangehörige, die in diesem Gebiet selbst wenig Ausbildung erfahren haben und das Thema möglicherweise gering schätzen, nicht erfolgreich sein kann im Hinblick auf Reliabilität und Validität der Beurteilungen. Diese pessimistische Ansicht läßt sich in dieser Studie nicht bestätigen. Im Gegenteil, insgesamt läßt sich feststellen, dass die Urteile von geschulten Prüfern mit den Urteilen von Experten auf dem Berliner Global Rating sehr gut übereinstimmen, insbesondere bei der Identifikation der leistungsschwachen und –starken Prüflinge.

Trainierte SPs geben ähnliche Bewertungen ab, stimmen jedoch weniger exakt mit den Experten überein. Die Bewertungen fallen tendenziell milder aus. Aufgrund der rechtlichen Situation in Deutschland ist es wenig wahrscheinlich, dass SP in bestehensrelevanten Prüfungen als Prüfer fungieren. Ein Einsatz in formativen, nicht-bestehensrelevanten Prüfungen ist aber durchaus denkbar und wird durch die Ergebnisse gestützt.

Die unjustierte Intraklassenkorrelation unterschätzt die Übereinstimmung eher, da Mittelwertunterschiede zwischen Ratern als Fehler einbezogen werden. Vor diesem Hintergrund sind die Ergebnisse sehr zufriedenstellend und können als guter Beleg für die Kriteriumsvalidität des Instruments bewertet werden. Die Zuverlässigkeit der Mittelwerte über mehrere Rater ist immer höher als die Zuverlässigkeit der Urteile eines einzelnen Raters. Vor dem Hintergrund, dass

[4] "It has often been argued that faculty cannot assess that which they do not see." Gray (1996), S. S57

Prüfer nicht ausgewählt werden können im Hinblick auf ihre Expertise im Bereich Gesprächsführung, ist die Durchführung eines Rater-Trainings besonders wichtig.

## 4.2 Konstruktvalidität

Um die Konstruktvalidität zu prüfen, wurden in der Studie zwei Strategien verfolgt. Zum einen beurteilte eine weitere Expertengruppe die Studierenden auf einem zweiten Instrument zur Prüfung kommunikativer Kompetenzen in der ärztlichen Gesprächsführung, der CCOG-Checkliste. Eine hohe Korrelation zwischen den mit unterschiedlichen Instrumenten gemessenen Ergebnissen wäre ein Beleg dafür, dass mit beiden Instrumenten das gleiche Konstrukt gemessen wird und damit für die konvergente Validität (Hypothese 3).

Zum anderen wurden die Prüfungsergebnisse der Studierenden in den OSCE-Stationen herangezogen, um die diskrimante Validität zu beurteilen. Es sollte kein allzu hoher Zusammenhang zwischen Prüfungsergebnis und kommunikativen Kompetenzen bestehen, da in der OSCE-Prüfung ein anderes Konstrukt - klinisch-praktische Fertigkeiten - gemessen wird als mit dem Berliner Global Rating (Fragestellung 4).

### 4.2.1 Konvergente Validität

Auf der CCOG-Checkliste wird beobachtetes Verhalten direkt notiert, während der Beurteiler bei der globalen Ratingskala seine Bewertung als Summe von Beobachtungen bilden muss. Eine hohe Übereinstimmung zwischen beiden Instrumenten würde dafür sprechen, dass mit dem Globalen Rating trotz augenscheinlich höherer Subjektivität das gleiche Konstrukt gemessen wird.

Die Daten zeigen hohe Korrelationen zwischen den von Experten beurteilten Leistungen der Studierenden auf dem Berliner Global Rating und auf der CCOG-Checkliste. Dies zeigt sich vor allem im Zusammenhang der Gesamtwerte über die Stationen in beiden Semestern. Diese Ergebnisse werden durch die hohe Ausprägung des Determinationskoeffizienten gestützt und sind ein Beleg für die konvergente Validität. Der Zusammenhang ist im 5. Semester deutlich stärker als im 3. Semester. Hier haben sich möglicherweise Trainingseffekte bemerkbar gemacht: Die Experten, die die CCOG-Checkliste verwendet haben, waren an beiden Prüfungstagen im Einsatz. Von einigen Experten wurde in einem Abschluß-Feedback zurückgemeldet, dass das

Ausfüllen der vielen Items in der begrenzten Zeit sehr anspruchsvoll war. Mit zunehmender Übung fiel dies möglicherweise am zweiten Prüfungstag (5. Semester) leichter. Dieser Erklärungsansatz wird durch die Tatsache gestützt, dass in den Checklisten-Daten des 5. Semesters weniger Werte fehlen. Möglicherweise konnten sich die Experten am zweiten Tag besser auf die Performanz der Studierenden konzentrieren, da sie mehr Übung im Beantworten der Items unter Zeitdruck hatten als am ersten Prüfungstag.

Falls dieser Erklärungsansatz zutreffend ist, spricht dies umso mehr für die Verwendung eines globalen Beurteilungsinstrument mit wenigen Items anstatt einer detaillierten Liste für die Beurteilung kommunikativer Kompetenzen in Prüfungssituationen, um Beurteilungsfehler bzw. fehlende Daten zu vermeiden, denn die Prüfer sind jeweils nur an einem Prüfungstag anwesend.

Die Betrachtung der deskriptiven Daten zeigt einen deutlich niedrigeren Mittelwert in Station 6 (Anamnese und Kniegelenksuntersuchung bei Rheuma) und eine höhere Zahl fehlender Werte, die die Anzahl der eingehenden Fälle in dieser Station auf n = 44 reduziert. Eine genauere Analyse zeigt, dass die Beurteilungen der beiden Rater sich erheblich unterscheiden: Die Studierenden erreichten bei einem Rater im Mittel 50,4% der Punkte (SD = 20,3%), während die Prüflinge beim anderen Rater im Schnitt nur 27,1% (SD = 7,6%) erreichten und damit durchgängig schlechter beurteilt wurden, wie auch die geringe Standardabweichung zeigt. In den mit dem globalen Beurteilungsinstrument erhobenen Daten schneiden die Prüflinge in dieser Station dagegen nicht schlechter ab. Die betroffene Expertin gab nach der OSCE-Prüfung in einer Abschluss-Feedbackrunde an, dass sie bei der Station „Anamnese und Kniegelenksuntersuchung bei Rheuma“ Schwierigkeiten bei der Beurteilung mit der CCOG-Checkliste hatte. Die zweite Beurteilerin hatte keine Schwierigkeiten beim Ausfüllen der Checkliste. Sicherheitshalber sollte daher bei Stationen, die auch eine Untersuchung beinhalten, sorgfältig ausgewählt werden, ob diese auch geeignet sind für die Prüfung der ärztlichen Gesprächsführung. Aufgrund der vorgegebenen Items einer detaillierten Checkliste ist das Einsatzspektrum wahrscheinlich stärker eingeschränkt, als es bei dem globalen Beurteilungsinstrument der Fall ist.

Die detaillierte Analyse der mit der CCOG-Checkliste erhobenen Daten liefert gleichzeitig eine wertvolle Rückmeldung für die Curriculum-Planung. Aufschlussreich ist neben den am häufigsten erfüllten Items die Auswertung der am seltensten erfüllten Items. Die Analyse der einzelnen Items der CCOG-Checkliste zeigte, dass weniger als 10% der Studierenden (Teil-)-Punkte bei vier Items erhalten hatten. Dies betraf die Items Nr. 5 („verdeutlichte Interviewablauf und Zeitrahmen“), Nr. 14 („explorierte Gefühle und Gedanken zum Problem“), Nr. 15 („ermutigte zum Ausdruck von Emotionen“) und Nr. 27 („schloß das Gespräch durch eine kurze

Zusammenfassung ab“). Diese Items können ein Hinweis auf Lücken im Curriculum oder auf Mängel in der Vermittlung im Unterricht sein. Es ist aber auch möglich, dass einzelne Aspekte der Checkliste in der Prüfungssituation von den Studierenden nicht als relevant eingestuft werden oder schwieriger umzusetzen sind als in der Lernsituation. Die Thematisierung von Gesprächsablauf und Zeitrahmen erscheint den Prüflingen möglicherweise eher vernachlässigbar, da diese weitgehend durch die Aufgabenstellung vorgegeben sind. Schwieriger umzusetzen ist wahrscheinlich die Zusammenfassung am Gesprächsende, da das Gespräch bei Ablauf der Prüfungszeit durch das Klingelsignal zum Teil abrupt beendet wird. Fehlende Punkte können daher sowohl auf mangelnde Gesprächsführung beim Konsultationsende zurückzuführen sein als auch durch mangelnde Effektivität in der Bewältigung der Aufgabe hinten übergefallen sein, also indirekt ein Problem der Kontextspezifität sein. Letztlich ist auch sorgfältig abzuwägen, ob einzelne Stationen möglicherweise weniger geeignet sind für die Prüfung kommunikativer Kompetenzen. Die Exploration von Gefühlen und die Ermutigung zum Emotionsausdruck spielt sicher eine unterschiedlich wichtige Rolle in Abhängigkeit von der Ausgestaltung der Patientenrolle in der jeweiligen Station.

Zusammenfassend kann zur Beantwortung von Hypothese 3 festgehalten werden, dass die hohen Zusammenhänge zwischen den Ergebnissen auf dem BGR und der CCOG-Checkliste einen Beleg für die konvergente Validität darstellen.

### 4.2.2 Diskriminante Validität

Sowohl kommunikative Kompetenzen als auch klinisch-praktische Fertigkeiten sind keine generellen Fähigkeiten, sondern hängen in gewissem Ausmaß vom Kontext ab. Es besteht immer eine Beziehung zwischen dem Vermögen einer Person und einer Aufgabe, die in einer bestimmten Situation bewältigt werden soll (89). Daher variiert die Performanz zwischen verschiedenen Aufgaben zum Teil erheblich. Dieser Sachverhalt ist in der Literatur als „Kontextspezifität“ bekannt. Da die Leistung in einer Station ein schlechter Prädiktor für die Performanz in einer anderen ist, wird empfohlen, eine genügend große Anzahl an Aufgaben in die Prüfung einzuschließen, um eine hohe Reliabilität sicherzustellen (65). Bei den kommunikativen Kompetenzen gilt ebenso, dass die Performanz zu einem Teil vom Inhalt abhängig ist: Fachwissen ist zu einem gewissen Grad Voraussetzung für eine gute ärztliche Gesprächsführung. Einem Prüfling, der mit einer gestellten Aufgabe hinsichtlich seines medizinischen Wissens und seiner klinisch-praktischen Fertigkeiten überfordert ist, wird es schwer fallen, ein gutes ärztliches

Gespräch zu führen. Umgekehrt kann sich ein Prüfling, dem die Bewältigung der Aufgabe leichtfällt, möglicherweise mehr auf die Gesprächsführung konzentrieren. Unklar ist bisher, ob alle Aspekte kommunikativer Kompetenzen im gleichen Ausmaß kontextspezifisch sind. Es gibt Hinweise, dass einige Fertigkeiten wie z. B. der Aufbau einer therapeutischen Beziehung weniger abhängig vom Inhalt sind (89).

Es ist also von einem gewissen Zusammenhang zwischen dem Ergebnis eines Studierenden auf der Checkliste in der OSCE-Station und seinem Ergebnis auf dem Berliner Global Rating auszugehen. Eine hohe oder sehr hohe Korrelation zwischen den Ergebnissen wäre jedoch ein Hinweis auf mangelnde diskriminante Validität und würde darauf hindeuten, dass (unbeabsichtigt) das gleiche Konstrukt gemessen wird. Um den Umfang dieser Beziehung abschätzen zu können und unbeabsichtigte Konsequenzen des Testens zu vermeiden, wurden die Prüfungsdaten der Studierenden im OSCE zur Analyse herangezogen.

Die Analyse zeigte in vier Stationen keine bis niedrige Zusammenhänge, in drei Stationen mittelmäßige und in einem Fall einen hohen Zusammenhang, wobei der Determinationskoeffizient lediglich 38% beträgt. Die Ergebnisse bewegen sich damit in einem Rahmen, der auch von anderen Autoren berichtet wird (z. B. (90)). Es läßt sich kein genereller Trend ablesen, dass beispielsweise Stationen mit Untersuchungen weniger Zusammenhang aufweisen als solche mit dem Schwerpunkt Anamneseerhebung.

Neben der Kontextspezifität ist ein weiterer Grund für Zusammenhänge darin zu suchen, dass die OSCE-Checklisten zum Teil auch kommunikationsspezifische Items enthielten. Das waren sowohl allgemein Items (z. B. zur Begrüßung und Vorstellung) als auch für die Aufgabenstellung wichtige Items. Letzteres betraf vor allem die Stationen 4 (Sexualanamnese), 8 (Anamnese Streßsyndrom) und 9 (Anamnese Panikattacke). Es ist daher davon auszugehen, dass die Zusammenhänge geringer ausfallen, wenn diese Überlappungen aus den Checklisten getilgt würden.

Zusammenfassend kann zur Beantwortung von Fragestellung 4 Folgendes festgehalten werden: Insgesamt ergeben sich Hinweise für die diskriminante Validität des Instruments, die durch Redundanzen zwischen BGR und Checklisten-Items eher unterschätzt werden dürfte. Mit dem Berliner Global Rating können die kommunikativen Kompetenzen der Studierenden ausreichend unbabhängig von den klinisch-praktischen Fertigkeiten erfaßt werden. Die Items der OSCE-Checklisten sollten bei gleichzeitiger Prüfung der kommunikativen Kompetenzen einer sorgfältigen Prüfung unterzogen werden. Bei Doppelungen muss entschieden werden, ob die Fertig-

keiten als so bedeutend angesehen werden, dass ein stärkeres Gewicht in der Benotung gewünscht wird oder ob ähnliche bzw. doppelte Items aus der Checkliste entfernt werden sollen. Die unterschiedliche Höhe der Zusammenhänge in den Stationen verweist auf die Wichtigkeit einer ausreichend großen Anzahl an Stationen, in denen kommunikative Kompetenzen geprüft werden sollten.

### 4.3 Rater Training

Das Rater-Training stellte sich als wirkungsvolles Instrument der Fakultätsentwicklung im Hinblick auf das Thema ärztliche Gesprächsführung heraus. Viele mündliche Rückmeldungen der Prüfer zeigten, dass das Training als Fortbildungsmaßnahme wertgeschätzt wurde. Einige Prüfer gaben an, aufgrund des Trainings dem Thema in Zukunft in der eigenen Arbeit und der Supervision jüngerer Kollegen mehr Aufmerksamkeit widmen zu wollen.

Während Frame-of-reference-Trainings die Genauigkeit der Raterurteile erhöhen, zielen sog. „Rater-Error-Trainings“ darauf ab, Beurteilungsfehler zu verringern. Zur Optimierung des Trainingseffekts könnte das im Rahmen der Studie entwickelte Frame-of-reference-Training noch um eine Rater-Error-Trainingskomponente ergänzt werden. Dies wäre hilfreich, um Milde- und Strengefehler zu minimieren.

Das Rater-Training kann sowohl unter dem Gesichtspunkt der Vermittlung der Handhabung des Berliner Global Rating als auch unter dem der Fakultätsentwicklung als sehr erfolgreich eingeschätzt werden. Gleichwohl ist es relativ aufwändig in der Durchführung. Hilfreich für den perspektivischen Einsatz des Berliner Global Rating in OSCE-Prüfungen des Reformstudiengangs und damit verbundenen notwendigen Trainingsmaßnahmen wäre die Unterstützung des Prüfungsausschusses bzw. der Fakultät. Für eine regelhafte Durchführung würde die Umsetzung sicher erleichtert, wenn die Teilnahme am Training vergütet werden könnte; dafür käme die leistungsorientierte Mittelvergabe oder die Zertifizierung durch die Ärztekammer in Betracht. Außerdem wäre es wichtig, ein oder zwei weitere Prüfer zu trainieren, falls Prüfer am Tag der Prüfung ausfallen oder aufgrund von nicht korrigierbaren Beurteilungsfehlern ausgeschlossen werden müssen. Im Hinblick auf die Fairness der Prüfung sollte kein Prüfer die Studierenden mit dem Berliner Global Rating beurteilen, der nicht an der Trainingsmaßnahme teilgenommen hat.

Im Weiteren stellt sich die Frage nach der Stabilität von Raterurteilen nach Trainingsmaßnahmen. In der Literatur findet dieses Thema kaum Beachtung. Übliche Zeiträume in Studien zur Retest-Reliabilität von wenigen Wochen sind kaum auf Prüfungszyklen übertragbar. Das gesamte Training jedes Mal zu wiederholen, scheint unrealistisch und würde die Prüfer frustrieren. Eine Möglichkeit bestünde darin, ein internetbasiertes Modul zur Wiederholung zu entwickeln oder für bereits geschulte Prüfer eine Kurzform des Trainings anzubieten. Um diese Frage zufriedenstellend beantworten zu können, sind weitere Studien notwendig.

## 4.4 Einschränkungen

Diese Studie wies verschiedene Limitationen auf: Zum einen konnte das geplante Design nicht vollständig umgesetzt werden, da vier der 20 Prüfer nicht am Rater-Training teilnehmen konnten. Drei dieser Personen sowie eine weitere hatten darüberhinaus schon einmal die Übung „Interaktion“ unterrichtet, so dass sie Quasi-Experten waren. Die Auswirkungen dieser Einschränkungen sind in Abschnitt 4.1.1 genauer analysiert.

Desweiteren war es mit diesem Studiendesign nicht möglich, Aussagen über Unterschiede zwischen den Semestern zu treffen, da die Studierenden in beiden Semester unterschiedliche Aufgaben zu bewältigen hatten. Es wäre interessant zu untersuchen, ob sich die von Hodges & McIlroy (80) berichteten Ergebnisse, dass sich Kompetenzunterschiede auf dem globalen Rating abbilden lassen, mit dem Berliner Global Rating replizieren lassen.

## 4.5 Abschließende Überlegungen zum Einsatz in klinisch-praktischen Prüfungen

Die Ergebnisse dieser Studie zeigen eine hohe exakte Übereinstimmung zwischen den Urteilen der Prüfer und der Experten, die in den Extremgruppen sogar sehr hoch ist. Die Übereinstimmung zwischen den Urteilen der Simulationspatienten und der Experten fiel geringer aus aufgrund milderer Urteile der Simulationspatienten, bewegt sich aber dabei auch noch auf hohem Niveau. Der Vergleich mit einem anderen Instrument zur Prüfung ärztlicher Gesprächsführung sowie mit den klinisch-praktischen Fertigkeiten der Prüflinge erbrachte solide Belege für die Konstruktvalidität des Berliner Global Rating.

Damit sind die Voraussetzungen für einen bestehensrelevanten Einsatz des Instruments zur Prüfung kommunikativer Kompetenzen im Rahmen von OSCE-Prüfungen gegeben. Es stellt sich die Frage, wie die Prüfung kommunikativer Kompetenzen in das Gesamt-Prüfungsgeschehen eingebettet werden sollte. Eine Prüfung kann eine Reihe verschiedener Funktionen haben: sie kann einen minimalen Kompetenzlevel sichern, Studierende in eine Rangreihenfolge bringen, Feedback zum Lernerfolg der Studierenden oder zum Unterricht leisten. Jeder Zweck erfordert unterschiedliche Standards und eine unterschiedliche Präsentation der Ergebnisse (57). Darüber hinaus hat eine Prüfung immer eine lernsteuernde Funktion. Durch die Tätigkeit der Prüfer bietet eine Prüfung immer auch eine Chance zur Fakultätsentwicklung.

Mit der Prüfung kommunikativer Kompetenzen mittels des Berliner Global Rating im Rahmen der OSCE-Prüfungen können unterschiedliche Funktionen erreicht werden:

Die Prüfung kommunikativer Kompetenzen im Rahmen der OSCE-Prüfung würde der Sicherung eines zumindest minimalen Kompetenzlevels in diesem Bereich dienen. Zu empfehlen wäre die Prüfung ab dem 3. Semester im Reformstudiengang, da die Studierenden ab diesem Semester in der Übung Interaktion mit Simulationspatienten üben. Das Berliner Global Rating scheint für alle Stationen geeignet, in denen Simulationspatienten eingesetzt werden (Anamnese- und Untersuchungsstationen oder Kombination). Bei Untersuchungsstationen ist es möglicherweise günstiger, wenn die Aufgabenstellung zumindest eine fokussierte Anamnese beinhaltet, weil dies die Beurteilung der kommunikativen Kompetenzen für die Prüfer erleichtert.

Die Beurteilung der kommunikativen Kompetenzen bei Medizinstudierenden sollte im Sinne hoher ökologischer Validität in einem ärztlichen Kontext stattfinden. Die isolierte Prüfung kommunikativer Kompetenzen ohne medizinischen Kontext führt zu einer hohen Künstlichkeit und gefährdet die Generalisierbarkeit der Ergebnisse im Hinblick auf das ärztliche Handeln.

Eine Integration in die OSCE-Stationen trägt auch der Kontextgebundenheit kommunikativer Kompetenzen Rechnung: Da kommunikative Kompetenzen nicht völlig unabhängig vom Kontext sind, sollte das Ergebnis auf dem Berliner Global Rating in einem festzulegenden Verhältnis mit dem Ergebnis der Checkliste für die klinisch-praktischen Fertigkeiten zu einem Stationsergebnis verrechnet werden und nicht als einzeln zu bestehender Prüfungsteil. Dies wäre ein Signal an die Studierenden, dass eine kompetente ärztliche Gesprächsführung nicht als „add-on“ betrachtet wird, sondern als integraler Bestandteil der Aufgabenstellung anzusehen ist. Die von dieser Festlegung ausgehende zu vermutende Lernsteuerung wäre wünschenswert. Schon im

Rahmen der Studie gab es verstärkt Interesse am Einsatz von zusätzlichen SPs zum Üben seitens einzelner Studierender, was die lernsteuernde Wirkung von Prüfungen zeigt.

Das Berliner Global Rating eignet sich gut als Ergänzung zu der Stations-Checkliste, die die klinisch-praktischen Fertigkeiten abbildet. Mittels des Instruments können so wichtige Aspekte der Performanz, die auf der bisherigen Checkliste nicht abgebildet sind, erfaßt werden. Eine häufige kritsche Rückmeldung von Prüfern war, dass es zur Zeit möglich ist, dass Studierende auf der Checkliste formal die gleiche Punktzahl erreichen, aber die Qualität der Performanz, gerade im Hinblick auf den Umgang mit den Simulationspatienten und die ärztliche Gesprächsführung, sehr unterschiedlich war. Die zusätzliche Prüfung der kommunikativen Kompetenzen würde es erlauben, hier stärker zu differenzieren und könnte daher eine wichtige Ergänzung in der OSCE-Prüfung darstellen. Damit wäre es möglich, die Studierenden in eine andere Rangreihenfolge zu bringen, die der tatsächlich gezeigten Leistung besser entspricht. Eventuelle inhaltliche Doppelungen müssen kritisch hinterfragt und ggf. die Stations-Checklisten modifiziert werden.

Aus testtheoretischer Perspektive ist die aufgrund zu kurzer Prüfungsdauer oft nicht hinreichende Reliabilität vieler OSCEs problematisch. In der Praxis muss dabei in Abhängigkeit von der Funktion der Prüfung (Feedback, bestehensrelevante Abschlussprüfung usw.) immer die Balance zwischen testtheoretischen Anforderungen und zur Verfügung stehenden Ressourcen sowie der Zumutbarkeit für die Prüflinge gefunden werden. Hinsichtlich der Reliabilität ist die Beurteilung der Gesprächsführung auf der Basis einer einzigen Situation sicherlich als unzureichend zu beurteilen. In der OSCE-Prüfung bietet sich jedoch die Möglichkeit, mehrere Aufgabenstellungen als Stichprobe des zu beurteilenden Merkmals heranzuziehen. Zudem hat die Prüfungsmethode OSCE den Vorteil, dass jeder Prüfling von mehreren, voneinander unabhängigen Prüfern beurteilt wird, wodurch sich eventuelle Urteilsfehler ausgleichen. Die Reliabilität läßt sich in größerem Maß durch die Erhöhung der Anzahl der Stationen steigern als durch die Anwesenheit mehrerer Prüfer in einer Station oder durch die Länge der einzelnen Stationen (91). Die Qualitätsverbesserung von Stationen bzw. der Ausschluss von weniger geeigneten Stationen sowie die Durchführung des Rater-Trainings können die Reliabilität ebenfalls erhöhen. Festzulegen wäre daher die mindestens notwendige Anzahl von Stationen. In einer Studie von Boulet et al. (92), die die Anforderungen an die Prüfung kommunikativer Kompetenzen hinsichtlich der Reliabilität untersuchte, wurde eine Anzahl von zehn Stationen als hinreichend für eine staatliche Zulassungsprüfung ermittelt. Dies ist sicher eine Herausforderung und für fakultätsinterne Prüfungen möglicherweise nicht realisierbar. Empirisch stützen ließen

sich Reliabilitätsüberlegungen durch die Analyse der Daten dieser Studie mittels der Generalisierbarkeitstheorie. Diese noch wenig genutzte Methode stellt eine Erweiterung der klassischen Testtheorie dar und erlaubt die Schätzung der Varianzkomponenten der Variablen, die in diesem Zusammenhang interessieren. In einer sogenannten „G-Study" lassen sich alle Fehlerquellen zugleich quantifizieren anstatt mehrere Koeffizienten unterschiedlicher Größe zu berechnen wie in der klassischen Testtheorie (66). Mittels einer Decision-Studie („D-study") ließe sich darüber hinaus ermitteln, wie viele Stationen aufgrund der Kontextspezifität notwendig wären, um eine ausreichende Reliabilität zu erreichen. Zukünftige G- bzw. D-Studien wären hilfreich, um Fragen zur Reliabilität zu klären.

Bei der Prüfung der kommunikativen Kompetenzen mittels des Berliner Global Rating kann für jeden Studierenden eine detaillierte Rückmeldung erstellt werden und die Prüfung so eine Feedback-Funktion erfüllen. Dies würde sicher auch zur Akzeptanz auf Seiten der Studierenden beitragen, die sich immer eine Rückmeldung über ihre Leistungen wünschen. Die detaillierten Ergebnisse der OSCE-Checklisten werden zur Zeit nicht veröffentlicht, sondern nur das Gesamtergebnis in der Station. Die Konstruktion der Stationen zu den Lernzielen der Semester ist ein arbeitsintensiver Prozeß, der die Erstellung von Aufgabenstellung, Patientenrolle und Checkliste sowie deren Überprüfung in einem „Trockendurchlauf" erfordert. Es wäre nicht realisierbar, jedes Jahr komplett neue Stationen zu entwickeln. Aus diesem Grund werden die Checklisten (bisher) geheimgehalten. Im Gegensatz zur Checkliste mit den klinisch-praktischen Fertigkeiten gibt es keinen Grund, das Berliner Global Rating als Beurteilungsinstrument den Studierenden nicht bekannt zu machen. Aus verschiedenen Gründen wird zur Zeit in der OSCE-Prüfungen auch auf mündliches Feedback verzichtet: zum einen verlängert sich dadurch die Prüfungsdauer, zum anderen könnte negatives Feedback die Leistung in der nächsten Station beeinträchtigen. Um das Feedback konkreter zu gestalten als es mit dem Berliner Global Rating möglich ist, wäre eine Möglichkeit, dass die Prüfer schriftliche Kommentare verfassen, die den Studierenden nach der Prüfung ausgehändigt werden könnten. Im Rater-Training könnte den Prüfern vermittelt werden, was beim Geben von (konstruktivem) Feedback zu beachten ist.

Das Rater-Training sollte unverzichtbare Voraussetzung für die Prüfung kommunikativer Kompetenzen mit dem Berliner Global Rating sein. Wie sich gezeigt hat, kann es auch einen Beitrag zur Fakultätsentwicklung leisten. Bezüglich der Stabilität der Rater-Urteile und der damit notwendigen Wiederholung des Trainings können noch keine Aussagen getroffen werden; hier wären weitere Untersuchungen notwendig und wünschenswert.

Mit dem Berliner Global Rating steht ein deutschsprachiges Instrument zur Verfügung, über dessen Qualität gesicherte Daten vorliegen und das in den zunehmend häufiger durchgeführten klinisch-praktischen Prüfungen in Deutschland eingesetzt werden kann. Mit der Beurteilung kommunikativer Kompetenzen der Studierenden als zentraler klinischer Fertigkeit nähme die Charité eine Vorreiterrolle in der medizinischen Ausbildung in Deutschland ein. Wenn kommunikativen Kompetenzen der Stellenwert zugeschrieben werden soll, der ihnen als Pflichtfach über zehn Semester im Reformstudiengang Medizin zugedacht ist, sollten sie auch geprüft werden. Das Einsatzgebiet des Berliner Global Rating ist nicht auf die Ausbildung oder die Medizin beschränkt: Das Instrument könnte auch Anwendung in der ärztlichen und psychologischen Weiterbildung und in anderen praktischen Prüfungsformen Verwendung finden wie dem Mini-CEX (clinical examination) oder der 360-Grad-Beobachtung. Voraussetzung wäre aber auch hier ein Rater-Training, um Beurteilungsfehler zu minimieren.

Van der Vleuten (93) schlägt vor, den Nutzen einer Prüfung anhand von fünf Kriterien zu beurteilen: Validität, Reliabilität, Lernsteuerung, Akzeptanz bei Studierenden und Fakultätsangehörigen sowie Kosten. Nicht alle diese Aspekte lassen sich auf der Grundlage der Studiendaten beurteilen; lernsteuernde Effekte oder Akzeptanz zeigen sich naturgemäß erst nach der Einführung einer neuen Prüfungsmethode. Die Prüfung kommunikativer Kompetenzen im Rahmen der OSCE-Prüfungen mittels Berliner Global Rating sollte im Verlauf unter diesen fünf Gesichtspunkten beurteilt werden. Hinsichtlich Validität, Lernsteuerung und Kosten kann schon jetzt ein positives Fazit gezogen werden.

Im Weiteren sollten die Stabilität der Urteile zur Abschätzung der notwendigen Frequenz des Rater-Trainings sowie die Reliabilität mittels Generalisierbarkeitsanalysen genauer untersucht werden.

# 5 Zusammenfassung

Das Gespräch mit Patienten ist eine häufige ärztliche Tätigkeit und ein wichtiges diagnostisches und therapeutisches Instrument. Eine patientenorientierte Gesprächsführung trägt nachweislich zu positiven Behandlungsergebnissen bei. Kommunikative und soziale Kompetenzen sind trainierbar, daher ist die Vermittlung im Rahmen des Medizinstudiums internationaler Standard. Zunehmend werden diese Kompetenzen auch geprüft. Für die Prüfung praktischer Fertigkeiten hat sich die Prüfungsmethode OSCE (objective structured clinical examination) etabliert: Alle Studierenden müssen identische Aufgaben in einem Prüfungsparcours bewältigen und werden von verschiedenen Prüfern dabei beurteilt. In einer OSCE-Prüfung kann in einer quasi-realen Situation ein großes Aufgabenspektrum geprüft werden. Dabei kommen häufig Simulationspatienten zum Einsatz, die Patientenfälle wiederholt standardisiert darstellen können.

Im Reformstudiengang Medizin der Charité werden die klinisch-praktischen Fertigkeiten der Studierenden nach fast jedem Semester mittels OSCE geprüft. Ärztliche Gesprächsführung in verschiedenen Kontexten wird über zehn Semester als Pflichtveranstaltung unterrichtet, bisher allerdings nicht (mit)geprüft.

Für die Prüfung kommunikativer Kompetenzen ist die Integration in die OSCE-Prüfung state of the art. Die Bewertung findet anhand detaillierter Checklisten oder globalerer Ratingskalen statt. Für den Bereich kommunikativer Kompetenzen wird der Einsatz globaler Instrumente empfohlen, die sich als mindestens gleichwertig hinsichtlich der psychometrischen Qualität erwiesen haben. Bislang stand jedoch kein adäquates deutschsprachiges Instrument zur Verfügung.

Gegenstand dieser Arbeit ist die Validierung einer kanadischen globalen Beurteilungsskala, die zuvor übersetzt und modifiziert wurde („Berliner Global Rating“). Dabei wurden verschiedene Strategien verfolgt: Zur Konstruktvalidierung (konkurrente Validität) wurde untersucht, inwieweit geschulte Prüfer bzw. Simulationspatienten mit Experten, deren Urteil als Referenzkriterium definiert wurde, in ihrem Urteil übereinstimmen. Um Aussagen zur Kriteriumsvalidität treffen zu können, wurden zwei Ansätze verfolgt: Eine weitere Gruppe von Experten beurteilte die Prüflinge auf einem anderen Instrument zur Prüfung ärztlicher Gesprächsführung, einer detaillierten Checkliste. Übereinstimmende Testergebnisse wären ein Beleg für die konvergente Validität. Weiterhin wurde anhand der OSCE-Prüfungsergebnisse untersucht, ob die kommunikativen Kompetenzen mit dem „Berliner Global Rating“ (BGR) ausreichend unabhängig von den klinisch-praktischen Fertigkeiten der Studierenden erfasst werden; dies diente zur Abschätzung der diskriminanten Validität.

Im Rahmen der Studie wurde ein zweistündiges Rater-Training entwickelt, um die Rater in der Beurteilung mit dem BGR zu schulen. Insgesamt wurden sechs Trainings durchgeführt. Die Studie fand im Rahmen der OSCE-Prüfungen des 3. und 5. Semesters Reformstudiengang im Februar 2004 statt. Insgesamt 113 Studierende wurden von 20 Prüfern, 22 Experten sowie 20 Simulationspatienten in acht verschiedenen OSCE-Stationen beurteilt.

Zwischen den Urteilen der Experten und der Prüfer zeigten sich hohe exakte Übereinstimmungen. Bei den 25% leistungsstärksten und –schwächsten Studierenden war die Übereinstimmung sogar sehr hoch. Die Übereinstimmungen zwischen Experten und Simulationspatienten waren etwas schwächer ausgeprägt, aber immer noch auf hohem Niveau. Diese Ergebnisse belegen die Kriteriumsvalidität des Berliner Global Rating.

Die Analyse der konvergenten Validität zeigte hohe Korrelationen mit dem weiteren Instrument, welches das gleiche Konstrukt - kommunikative Kompetenzen - messen soll. Beim Vergleich der Studiendaten mit den Ergebnissen der Studierenden in der OSCE-Prüfung (diskriminante Validität) zeigten sich überwiegend keine bis mäßige Zusammenhänge zwischen kommunikativen Kompetenzen und den klinisch-praktischen Fertigkeiten. Beide Ergebnisse sind solide Belege für die Konstruktvalidität des Berliner Global Rating.

Die Ergebnisse zeigen, dass die kommunikativen Kompetenzen von Medizinstudierenden mit dem „Berliner Global Rating“ im Rahmen von OSCE-Prüfungen valide erfaßt werden können. Damit steht ein deutschsprachiges Instrument zur Verfügung, über dessen Qualität gesicherte Daten vorliegen und das in den zunehmend häufiger durchgeführten klinisch-praktischen Prüfungen in Deutschland eingesetzt werden kann. Im Rahmen der Studie hat sich das Instrument als praktikabel erwiesen. Die Implementation in OSCE-Prüfungen ließe sich damit annähernd kostenneutral umsetzen. Die bestehensrelevante Prüfung kommunikativer Kompetenzen hätte eine lernsteuernde Wirkung sowie eine Feedback-Funktion für die Studierenden; gleichzeitig würde sichergestellt, dass die Studierenden über ausreichende Kompetenzen in diesem Bereich verfügen.

Um die Fairness der Beurteilung zu gewährleisten und Urteilsfehler zu minimieren, ist ein Rater-Training für die Beurteiler essentiell. Mit dem im Rahmen der Studie konzipierten Rater-Training wurden gute Erfahrungen gemacht.

Im Weiteren sollten die Stabilität der Urteile zur Abschätzung der notwendigen Frequenz des Rater-Trainings sowie die Reliabilität mittels Generalisierbarkeitsanalysen genauer untersucht werden.

# 6 Literaturverzeichnis


1. Kurtz S, Silverman J, Draper J. Teaching and Learning Communication Skills in Medicine. Oxon, England: Radcliffe Medical Press Ltd.; 1998.

2. Simpson M, Buckman R, Stewart M, et al. Doctor-patient communication: the Toronto consensus statement. BMJ 1991;303:1385-7.

3. Buddeberg C, ed. Psychosoziale Medizin. 3rd ed. Berlin: Springer; 2004.

4. De Valck C, Van de Woestijne KP. Communication problems on an oncology ward. Patient Educ Couns 1996;29(2):131-6.

5. Coulter A, Magee H, eds. The European Patient of the Future. Maidenhead: Open University Press; 2003.

6. Bahrs O. Mein Hausarzt hat Zeit für mich - Wunsch und Wirklichkeit. Ergebnisse einer europäischen Gemeinschaftsstudie. G + G Wissenschaft 2003;3(1):17-23.

7. Tamblyn R, Abrahamowicz M, Dauphinee D, et al. Physician Scores on a National Clinical Skills Examination as Predictors of Complaints to Medical Regulatory Authorities. JAMA 2007;298(9):993-1001.

8. Evans BJ, Stanley RO, Burrows GD. Communication Skills Training and Patients' Satisfaction. Health Commun 1992;4(2):155-70.

9. Margalit APA, Glick, S.M., Benbassat,J., Cohen, A. Effect of a Biopsychosocial Approach on Patient Satisfaction and Patterns of Care. J Gen Intern Med 2004;19:485-91.

10. Roter DL, Hall JA, Kern DE, et al. Improving physicians' interviewing skills and reducing patients' emotional distress. A randomized clinical trial. Arch Intern Med 1995;155(17):1877-84.

11. Maguire P, Fairbairn, S., Fletcher, C. Consultation skills of young doctors: Benefits of feedback training in interviewing as students persist. BMJ 1986;292:1573-8.

12. Stewart MA, Brown JB, Donner A, et al. The Impact of Patient-Centered Care on Outcomes. JFP 2000;49(9):796-804.

13. Mead N, Bower P. Patient-centredness: a conceptual framework and review of the empirical literature. Soc Sci Med 2000;51(7):1087-110.

14. Platt FW, Gaspar DL, Coulehan JL, et al. "Tell Me about Yourself": The Patient-Centered Interview. Ann Intern Med 2001;134(11):1079-85.

15. Stewart MA. Effective physician-patient communication and health outcomes: A review of the literature. Can Med Assoc J 1995;152(9):1423-33.

16. Hargie O, Dickson D, Boohan M, et al. A survey of communication skills training in UK Schools of Medicine: present practices and prospective proposals. Med Educ 1998;32(1):25-34.

17. Aspegren K. BEME Guide No. 2: Teaching and learning communication skills in medicine - a review with quality grading of articles. Med Teach 1999;21(6):563 - 70.

18. Langewitz W, Eich P, Kiss A, et al. Improving communication skills-a randomized controlled behaviorally oriented intervention study for residents in internal medicine. Psychosom Med 1998;60(3):268-76.

19. Yedidia MJ, Gillespie CC, Kachur E, et al. Effect of Communications Training on Medical Student Performance. JAMA 2003;290(9):1157-65.

20. Bloom BS. Effects of continuing medical education on improving physician clinical care and patient health: A review of systematic reviews. Intl J of Technology Assesment in Health Care 2005;21(3):380-5.

21. Maguire P, Pitceathly C. Key communication skills and how to acquire them. BMJ 2002;325:697-700.

22. Barrows H, Abrahamson S. The Programmed Patient: a Technique for Appraising Student Performance in Clinical Neurology. J Med Educ 1964;39:802-5.

23. Collins J, Harden R. AMEE Medical Education Guide No. 13: real patients, simulated patients and simulators in clinical examinations. Med Teach 1998;20(6):508 - 21.

24. Duffy DF, Gordon GH, Whelan G, et al. Assessing Competence in Communication and Interpersonal Skills: The Kalamazoo II Report. Acad Med 2004;79(6):495-507.

25. General Medical Council (GMC). Tomorrow's doctors: Recommendations on undergraduate medical education. London: General Medical Council; 1993.

26. General Medical Council (GMC). Tomorrow's doctors: Recommendations on undergraduate medical education. 2003. (Accessed December 8, 2008, at http://www.gmc-uk.org/education/undergraduate/GMC_tomorrows_doctors.pdf.)

27. Liaison Committee on Medical Education (LCME). Functions and structure of a medical school. Standards for Accreditation of Medical Education Ptograms Leading to the M.D. Degree. 2007. (Accessed December 8, 2008, at http://www.lcme.org/functions2008jun.pdf.)

28. Accreditation Council for Graduate Medical Education (ACGME). Toolbox of Assessment Methods. 2000. (Accessed December 8, 2008, at http://www.acgme.org/Outcome/assess/Toolbox.pdf.)

29. Royal College of Physicians and Surgeons of Canada (RCPSC). CanMEDs Framework. 1996. (Accessed June 02, 2008, at http://rcpsc.medical.org/canmeds/about_e.php.)

30. Royal College of Physicians and Surgeons of Canada (RCPSC). CanMEDs 2000: Extract from the Can MEDs 2000 Project Societal Needs Working Group Report. Med Teach 2000;22(6):549-54.

31. Makoul G. Essential elements of communication in medical encounters: the Kalamazoo consensus statement. Acad Med 2001;76:390-3.

32. van Dalen J, Bartholomeus P, Kerkhofs E, et al. Teaching and assessing communication skills in Maastricht: the first twenty years. Med Teach 2001;23(3):245 - 51.

33. Hofer M, Jansen M, Soboll S. Verbesserungspotenzial des Medizinstudiums aus retrospektiver Sicht von Facharztprüflingen. Dtsch Med Wochenschr 2006(8):373-8.

34. Hemmer-Schanze C, Fueßl, H. S. Gesundheitsfaktor Zuhören. Munch Med Wochenschr 2006;148(1):1-8.

35. Jungbauer J, Kamenik C, Alfermann D, et al. Wie bewerten angehende Ärzte rückblickend ihr Medizinstudium? Ergebnisse einer Absolventenbefragung. Gesundheitswesen 2004;66:51-6.

36. bvmd. Kerncurriculum für die medizinische Ausbildung in Deutschland - Ein Vorschlag der Medizinstudierenden Deutschlands. 2006. (Accessed December 8, 2008, at http://bvmd.de/fileadmin/SCOME/Kerncurriculum/Kerncurriculum.pdf.)

37. Kiessling C, Dieterich A, Fabry G, et al. Basler Consensus Statement "Kommunikative und soziale Kompetenzen im Medizinstudium": Ein Positionspapier des GMA-Ausschusses Kommunikative und soziale Kompetenzen. GMS Z Med Ausbild 2008;25(2):Doc83.

38. Bundesministerium für Gesundheit. Approbationsordnung für Ärzte (ÄAppO) vom 27.06.2002. BGBl Teil 1 2002:2405-35.

39. Köhle K, Koerfer A, Thomas W, et al. Integrierte Psychosomatik: Beiträge zu einer Reform des Medizinstudiums. Psychother Psych Med 2003;53:65-70.

40. Sennekamp M, Gilbert K, Gerlach F. Anamneseerhebung und Gesprächsführung. In: Poster präsentiert auf der Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA) November 16-18, 2007; Hannover.

41. Kopecky-Wenzel M, Maier E, Muntau A, et al. Überbringen schlechter Nachrichten - videogestützte Trainingseinheit für Medizinstudenten. In: Poster präsentiert auf der Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA) November 16-18, 2007; Hannover.

42. Aders Y, Schafer T, Rusche H. The development of interactional competence in the medical reform curriculum at Ruhr-University Bochum. In: Poster presented at the Conference of the Association for Medical Education in Europe (AMEE) September 5-8, 2004; Edinburgh, Großbritannien.

43. Petersen C, Busche W, Bergelt C, et al. Kommunikationstraining als Teil des Medizinstudiums: ein Modellversuch. GMS Z Med Ausbild 2005;22(1):Doc08.

44. Allert G, Gommel M, Tamulionyté L, et al. Das interdisziplinäre Längsschnittcurriculum "Medizinische Psychologie, Psychotherapie und Psychosomatik" an der Universität Ulm. Psychother Psych Med 2002(8):355-62.

45. Schmitt GM, Kammerer E, Holtmann M. Förderung interaktioneller Kompetenzen von Medizinstudierenden. Psychother Psych Med 2003(9/10):390-8.

46. Jünger J, Köllner V. Integration eines Kommunikationstrainings in die klinische Lehre. Psychother Psych Med 2003;53:56-64.

47. Schildmann J, Härlein J, Burchardi N, et al. Die Aufklärung schwer kranker Patienten im interprofessionellen Kontext. GMS Z Med Ausbild 2006;23(4):Doc67.

48. Kampmann M, Schwantes U. "Patientenzentrierte Medizin" in der Ausbildung. JKM 2005;42:109-21.

49. Fischer T, Simmenroth-Nayda A, Herrmann-Lingen C, et al. Medizinische Basisfähigkeiten - ein Unterrichtskonzept im Rahmen der neuen Approbationsordnung. Z Allgemeinmed 2003;79(9):432-6.

50. Decker O, Rockenbauch K. Das neue Lehrkonzept für Medizinische Psychologie und Medizinische Soziologie an der Universität Leipzig - erste Eindrücke und Evaluation. Zeitschrift fuer Medizinische Psychologie 2006;15(1):27-30.

51. Burger W, Frömmel C. Der Berliner Reformstudiengang Medizin. Zielsetzungen und erste Erfahrungen. Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz 2002;45:152-8.

52. Kurtz SM, Silverman JD. The Calgary-Cambridge Observation Guides: an aid to defining the curriculum and organizing the teaching in commuication training programmes. Med Educ 1996;30:83-9.

53. Terzioglu P, Jonitz B, Schwantes U, et al. Kommunikative und soziale Kompetenzen. Vermittlung muss im Medizinstudium beginnen. Dtsch Ärztebl 2003;100:A 2277-9.

54. Mühlinghaus I, Fröhmel A. Kommunikative Kompetenzen. In: Voderholzer U, ed. Lehre im Fach Psychiatrie und Psychotherapie - ein Handbuch. Stuttgart: Kohlhammer Verlag; 2007:185-204.

55. Humphris GM, Kaney S. The Liverpool Brief Assessment System for Communication Skills in the Making of Doctors. Adv Health Sci Educ Theory Pract 2001;6(1):69-80.

56. Cate TJt. Summative assessment of medical students in the affective domain. Med Teach 2000;22(1):40-3.

57. Crossley J, Humphris G, Jolly B. Assessing health professionals. Med Educ 2002;36(9):800-4.

58. van der Vleuten CPM, Dolmans DHJM, Scherpbier AJJA. The need for evidence in education. Med Teach 2000;22(3):246-50.

59. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Med Educ 2004;38(2):199-203.

60. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65(9):S63-7.

61. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach 2007;29(9):855 - 71.

62. Harden R, Gleeson F. Assessment of clinical comptence using an objective structured clinical examination (OSCE). Med Educ 1979;13(1):41-54.

63. Amelang M. Psychologische Diagnostik und Intervention. 3rd ed. Berlin: Springer; 2002.

64. Bortz J, Döring N. Forschungsmethoden und Evaluation. 2nd ed. Berlin: Springer-Verlag; 1995.

65. van der Vleuten CPM, Norman G, de Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. Med Educ 1991;25:110-8.

66. Crossley J, Davies H, Humphris G, et al. Generalisability: a key to unlock professional assessment. Med Educ 2002;36(10):972-8.

67. Wass V, van der Vleuten C, Shatzer J, et al. Assessment of clinical competence. The Lancet 2001;357:945-49.

68. Fröhmel A, Burger W, Ortwein H. Einbindung von Simulationspatienten in das Studium der Humanmedizin in Deutschland. Dtsch Med Wochenschr 2007;132(11):549-54.

69. Georg W, Schubert S, Scheffner D, et al. Fünf Jahre Prüfungen im Reformstudiengang Medizin an der Charité-Universitätsmedizin Berlin. GMS Z Med Ausbild 2006;23(3):Doc48.

70. Cushing A. Assessment of Non-Cognitive Factors. In: Norman GR, ed. International Handbook of Research in Medical Education. Dordrecht: Kluwer Academic Publishers; 2002:711-55.

71. Hodges B, Regehr G, McNaughton N, et al. OSCE checklists do not capture increasing levels of expertise. Acad Med 1999;74(10):1129-34.

72. Swartz MH, Colliver JA, Bardes CL, et al. Global Ratings of Videotaped Performance versus Global Ratings of Actions Recorded on Checklists: A Criterion for Performance Assessment with Standardized Patients. Acad Med 1999;74(9):10281032.

73. Regehr G, MacRae H, Reznick R, et al. Comparing the Psychometric Properties of Checklists and Global Rating Scales for Assessing Performance in an OSCE-format Examination. Acad Med 1998;73(9):993-7.

74. van Dalen J, Prince CJAH, Scherpbier AJ, et al. Evaluating Communication Skills. Adv Health Sci Educ Theory Pract 1998;3:187-95.

75. Boon H, Stewart M. Patient-physician communication assessment instruments: 1986 to 1996 in review. Patient Educ Couns 1998;35(3):161-76.

76. Miller S, Hope T, Talbot D. The development of a structured rating schedule (the BAS) to assess skills in breaking bad news. Br J Cancer 1999;80(5/6):792-800.

77. Lang F, McCord R, Harvill L, et al. Communication assessment using the common ground instrument: psychometric properties. Fam Med 2004;36(3):189-98.

78. McLeod PJ, Tamblyn R, Benaroya S, et al. Faculty ratings of resident humanism predict patient satisfaction ratings in ambulatory medical clinics. J Gen Intern Med 1994;9(6):321-6.

79. Schnabl G, Hassard T, Kopelow M. The assessment of interpersonal skills using standardized patients. Acad Med 1991;66(9):S34-S6.

80. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ 2003;37(11):1012-6.

81. Wirtz M, Caspar F. Beurteiler-Übereinstimmung und Beurteiler-Reliabilität. Göttingen: Hogrefe; 2002.

82. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: A quantitative review. JOOP 1994;67:189-205.

83. Williams RG, Klamen DA, McGaghie WC. Cognitive, Social and Environmental Sources of Bias in Clinical Performance Ratings. Teach Learn Med 2003;15(4):270-92.

84. Lin LI-K. A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989;45:255-68.

85. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8(2):135-60.

86. Solomon DJ, Szauter K, Rosebraugh CJ, et al. Global Ratings of Student Performance in a Standardized Patient Examination: Is the Whole More than the Sum of the Parts? Adv Health Sci Educ Theory Pract 2000;5(2):131-40.

87. Heine N, Garman K, Wallace P, et al. An analysis of standardised patient checklist errors and their effect on student scores. Med Educ 2003;37(2):99-104.

88. Gray JD. Global rating scales in residency education. Acad Med 1996;71(1):S55-63.

89. Epstein RM. Medical Education: Assessment in Medical Education. NEJM 2007;356(4):387-96.

90. Hodges B, Turnbull J, Cohen R, et al. Evaluating communication skills in the OSCE format: reliability and generalizability. Med Educ 1996;30(1):38-43.

91. Newble D, Swanson DB. Psychometric characteristics of the objective structured clinical examination. Med Educ 1988;22:325-34.

92. Boulet J, Friedman Ben-David M, Ziv A, et al. High-Stakes Examinations: What do we know about measurement? Acad Med 1998;73(10):S94-S6.

93. van der Vleuten CPM. The Assessment of Professional Competence: Developments, Research and Practical Implications. Adv Health Sci Educ Theory Pract 1996;1(1):41-67.

# 7 Anhang

## 7.1 Globales Beurteilungsinstrument (englische Originalfassung)

### Appendix 1 Analytic global rating form used in the OSCE

| | | | | |
|---|---|---|---|---|
| **Response to patient's feelings and needs (empathy)** | | | | |
| 1 | 2 | 3 | 4 | 5 |
| Does not respond to obvious patient cues (verbal and non-verbal) and/or responds inappropriately | | Responds to patient's needs and cues, but not always effectively | | Responds consistently in a perceptive and genuine manner to the patient's needs and cues |
| **Degree of coherence in the interview** | | | | |
| 1 | 2 | 3 | 4 | 5 |
| No recognisable plan to the interaction; the plan does not demonstrate cohesion or the patient must determine the direction of the interview | | Organisational approach is formulaic and minimally flexible and/or control of the interview is inconsistent | | Superior organisation, demonstrating command of cohesive devises, flexibility, and consistent control of the interview |
| **Verbal expression** | | | | |
| 1 | 2 | 3 | 4 | 5 |
| Communicates in manner that interferes with and/or prevents understanding by patient, or communicates inappropriately with the patient | | Exhibits sufficient control of expression to be understood by an active, engaged listener (patient) | | Exhibits command of expression (fluency, diction, grammar, vocabulary, tone, volume and modulation of voice, rate of speech, pace and pronunciation |
| **Non-verbal expression** | | | | |
| 1 | 2 | 3 | 4 | 5 |
| Fails to engage, frustrates and/or antagonises the patient | | Exhibits enough control of non-verbal expression to engage a patient willing to overlook deficiencies such as passivity, self-consciousness or inappropriate aggressiveness | | Exhibits finesse and command of non-verbal expression (eye-contact, gesture, posture, use of silence, etc.) |
| **Overall assessment of the knowledge and skills demonstrated in the interview** | | | | |
| A = Incompetent | | B = Borderline | | C = Competent |
| Responds inappropriately and ineffectively to the task, indicating a lack of knowledge and/or undeveloped interpersonal and interviewing skills | | Responds effectively to some components of the task, some development of interpersonal and interviewing skills | | Responds precisely and perceptively to the task, consistently integrating all components |

Developed in collaboration with Cleo Boyd, Communications Consultant, Faculty of Pharmacy, University of Toronto.

Quelle: Hodges & McIlroy, 2003

## 7.2 Berliner Global Rating (deutsche Adaption)

| **Eingehen auf die Gefühle und Bedürfnisse der Patientin (Empathie):** | | |
|---|---|---|
| Die Studierende[1] geht durchgehend verständnisvoll auf die (verbalen und nonverbalen) Hinweise und Bedürfnisse der Patientin ein<br>**oder**<br>sie reagiert angemessen. | 1 2 3 4 5 | Die Studierende geht nicht auf offensichtliche (verbale und nonverbale) Hinweise und Bedürfnisse der Patientin ein<br>**oder**<br>sie reagiert unangemessen. |

| **Logischer Zusammenhang des Gesprächs (Struktur):** | | |
|---|---|---|
| Das Gespräch ist hervorragend organisiert;<br>das Vorgehen zeigt, dass die Studierende in der Lage ist, das Gespräch zusammenhängend zu gestalten und die Gesprächsführung in der Hand zu behalten. | 1 2 3 4 5 | Das Gespräch ist nicht erkennbar organisiert;<br>das Vorgehen wirkt zusammenhanglos oder die Patientin muss den Verlauf des Gesprächs festlegen. |

| **Verbaler Ausdruck:** | | |
|---|---|---|
| Die Studierende kommuniziert in einer Art und Weise, die es der Patientin leicht macht, sie zu verstehen<br>**oder**<br>sie kommuniziert angemessen mit der Patientin (z. B. hinsichtlich Wortwahl, Grammatik, Intonation, Lautstärke, Stimmmodulation, Sprechtempo und Aussprache). | 1 2 3 4 5 | Die Studierende kommuniziert in einer Art und Weise, die es der Patientin erschwert oder unmöglich macht, sie zu verstehen<br>**oder**<br>sie kommuniziert unangemessen mit der Patientin (z. B. hinsichtlich Wortwahl, Grammatik, Intonation, Lautstärke, Stimmmodulation, Sprechtempo und Aussprache). |

| **Nonverbaler Ausdruck:** | | |
|---|---|---|
| Die Studierende bezieht die Patientin durch nonverbalen Ausdruck durchgängig ein<br>**oder**<br>motiviert sie zur Gesprächsbeteiligung (z. B. durch Augenkontakt, Mimik, Gestik, Körperhaltung und Einsatz von Pausen). | 1 2 3 4 5 | Die Studierende schafft es nicht, die Patientin durch nonverbalen Ausdruck einzubeziehen und frustriert sie<br>**oder**<br>bringt sie gegen sich auf (z. B. durch Augenkontakt, Mimik, Gestik, Körperhaltung und Einsatz von Pausen). |

[1] Der besseren Lesbarkeit wegen wird im Text nur die weiblichen Endung genannt. Gemeint sind natürlich immer beide Geschlechter.

### 7.3 Calgary-Cambridge Observation Guide (CCOG)

Deutsche Kurzfassung (übersetzt von Dr. med. Heiderose Ortwein)

| | Ja | Teils | Nein | Trifft nicht zu |
|---|---|---|---|---|
| **Gesprächsbeginn**<br>1. Begrüßte den Patienten | | | | |
| 2. Stellte sich mit Namen vor | | | | |
| 3. Zeigte Respekt | | | | |
| 4. Identifizierte Probleme und Beschwerden (und vergewisserte sich, alles erfasst zu haben) | | | | |
| 5. Verdeutlichte Interviewablauf und Zeitrahmen | | | | |
| **Informationsgewinn**<br><u>Problemexploration</u><br>6. Ermutigte zum Erzählen | | | | |
| 7. Wechselte angemessen von offenen zu geschlossenen Fragen | | | | |
| 8. Hörte aufmerksam zu | | | | |
| 9. Unterstützte Patientenäußerungen verbal oder non-verbal | | | | |
| 10. Verwendete eine einfach verständliche Ausdrucksweise | | | | |
| 11. Klärte (unklare) Patientenäußerungen | | | | |
| 12. Gewann Daten für die aktuelle Krankheitsanamnese | | | | |
| <u>Verständnis für die Patientenperspektive</u><br>13. Stellte die Ideen des Patienten bzgl. der Ursache fest und respektierte diese | | | | |
| 14. Explorierte die Gefühle und Gedanken zum Problem | | | | |
| 15. Ermutigte zum Ausdruck von Emotionen | | | | |
| 16. Griff verbale und/oder non-verbale Signale auf | | | | |
| <u>Struktur der Konsultation</u><br>17. Fasste Gesagtes/Äußerungen vor einem Themenwechsel zusammen | | | | |
| 18. Fuhr im Gespräch fort indem er Überleitungen nutzte | | | | |
| 19. Strukturierte das Gespräch logisch | | | | |
| 20. Hielt den Zeitrahmen ein | | | | |
| <u>Beziehungsaufbau</u><br>21. Zeigte angemessenes non-verbales Verhalten | | | | |
| 22. Wenn er las, beeinträchtigte dies nicht den Dialog und die Beziehung zum Patienten | | | | |
| 23.War nicht wertend | | | | |
| 24. War verständnisvoll und unterstützend dem Patienten gegenüber | | | | |
| 25. War sicher/souverän | | | | |
| <u>Konsultationsende</u><br>26. Ermunterte den Patienten zusätzliche/offene Punkte anzusprechen | | | | |
| 27. Schloss das Gespräch durch eine kurze Zusammenfassung ab | | | | |
| 28. Vereinbarte mit dem Patienten die nächsten Behandlungsschritte | | | | |

# 8 Danksagung

Mein herzlicher Dank gilt meinem Betreuer Prof. Dr. Walter Burger (DRK-Klinik Berlin Westend) für die Unterstützung dieser Arbeit und seine wertvollen Anregungen. Bei der Erstellung dieser Arbeit haben mich außerdem Dipl.-Päd. Waltraud Georg (Ärztin, Assessment-Bereich, Charité) und Dipl.-Psych. Isabel Mühlinghaus (Reformstudiengang Medizin, Charité) durch ihre konstruktive Kritik sehr unterstützt, ebenso wie Dr. Edith Braun (FB Erziehungswissenschaft und Psychologie, FU Berlin).

An der Planung und Durchführung der Studie waren neben Isabel Mühlinghaus auch Annette Fröhmel (Ärztin, ehemals Reformstudiengang Medizin) und Dr. Heiderose Ortwein (Klinik für Anästhesie und operative Intensivmedizin, Charité) beteiligt; ihnen gilt mein Dank.

Waltraud Georg und Sebastian Schubert (Arzt, Assessment-Bereich) haben die Umsetzung des Forschungsprojekts tatkräftig unterstützt.

PD Dr. Dr. Werner Hopfenmüller (Institut für Biometrie und klinische Epidemiologie, Charité) gab mir Anregungen zur Auswertung der Daten.

Ich danke Dr. Brian Hodges für die Erlaubnis, das in Toronto entwickelte Instrument zu verwenden und für sein Interesse an dieser Studie.

Mein besonderer Dank gilt R. B. für seine uneingeschränkte Unterstützung.

# 9 Lebenslauf

Mein Lebenslauf wird aus Datenschutzgründen in der elektronischen Version meiner Arbeit nicht mit veröffentlicht.

# 10 Erklärung

„Ich, Simone Scheffer, erkläre, dass ich die vorgelegte Dissertationsschrift mit dem Thema: *Validierung des „Berliner Global Rating“ (BGR) - ein Instrument zur Prüfung kommunikativer Kompetenzen Medizinstudierender im Rahmen klinisch-praktischer Prüfungen (OSCE)* selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, ohne die (unzulässige) Hilfe Dritter verfasst und auch in Teilen keine Kopien anderer Arbeiten dargestellt habe.“

Berlin, den 24.01.2009