

4 Diskussion

Die Bereitschaft der Studierenden, an der zusätzlichen OSCE-Station teilzunehmen, war äußerst hoch. Insgesamt nahmen 94,8% aller OSCE-Prüflinge an der Kommunikationsstation teil. Im dritten Semester nahmen bis auf eine Studentin alle OSCE-Teilnehmer und im fünften Semester alle Studierenden teil. Im ersten Semester lag die Teilnehmerquote immerhin bei 87% der OSCE-Teilnehmer. Ein Grund für die hohe Bereitschaft zur Teilnahme ist sicherlich der „Pilotcharakter“ des Reformstudienganges. Die Studierenden werden regelmäßig während des Studiums mit verschiedenen Untersuchungen bezüglich medizinischer Ausbildung konfrontiert. Eine wichtige Voraussetzung für die hohen Teilnehmerquoten war dennoch die sorgfältige Integration der Kommunikationsstation in den bestehensrelevanten und teilnahmeverpflichtenden Teil des OSCE. Die Studierenden sollten bei der Teilnahme an der zusätzlichen OSCE-Station keine Beeinflussung ihrer Leistung in den restlichen bestehensrelevanten OSCE-Stationen befürchten. Frühzeitige Informationen über die Studie im Rahmen der Prüfungsanmeldung beugten dem vor. Gründe für die höheren Teilnehmerquoten im dritten und fünften Semester sind vermutlich der gewohnte Umgang mit OSCE-Prüfungen, bereits gemachte Erfahrung in Simulationspatientengesprächen, die ab dem zweiten Semester im Rahmen der Interaktions-Übung regelmäßig durchgeführt werden und der gewohnte Patientenkontakt im Rahmen des wöchentlich stattfindenden Praxistages. Die Studierenden des ersten Semesters absolvierten dagegen ihren ersten OSCE und konnten im ersten Semester noch keine Erfahrungen in Simulationspatientengesprächen bzw. am Praxistag sammeln.

Neben der hohen Teilnehmerquote deutet auch die niedrige Abbrecherquote in allen drei Semestern auf eine gute Integration der Kommunikationsstation in den OSCE-Ablauf hin. Die Studierenden wurden mit der Aufgabenstellung nicht überfordert. Besonders bei den Studierenden des ersten Semesters mit fehlender OSCE-Erfahrung und fehlendem SP-Kontakt hätten mehr als drei „Abbrecher“ erwartet werden können.

Vor der Diskussion der Frage, inwieweit sich die Ergebnisse der Studierenden in Abhängigkeit vom Semester unterscheiden, werden Mittelwerte und Standardabweichungen der Checkliste bzw. des Global Rating betrachtet.

4.1 Ergebnisse der Checkliste und des Global Rating

Die Mittelwerte der Prüfer- und SP-Bewertung (s. Tab. 5 - 7) liegen bei fast allen Items in einem Wertebereich zwischen 3 (akzeptabel) und 4 (gut). Lediglich das Item „Lautstärke und Betonung“ und die Items bezüglich des Gesprächsendes der Prüferbewertung fallen mit besonders hohen bzw. besonders niedrigen Mittelwerten auf. (s. Abb. 7) Das Item „Lautstärke

und Betonung“ scheint dem Großteil der Studierenden keine Schwierigkeiten bereitet zu haben. Die Mittelwerte liegen im Bereich von 4,1 im ersten Semester bis 4,5 im fünften Semester zwischen gut und hervorragend. Im Gegensatz dazu fallen deutlich schlechtere Bewertungen für den Bereich *Gesprächsende* auf. Hier ist besonders das Item „Klärung noch offener Fragen“ betroffen, mit einem Gesamtmittelwert von 1,69 zwischen inakzeptabel und schwach. Eine Ursache könnte sein, dass der Zeitrahmen von acht Minuten für das Gespräch mit der Simulationspatientin zu kurz war und das Gesprächsende darunter „gelitten“ hat. Um den Studierenden eine zeitliche Orientierung zu geben, war die Zwei-Minuten-Regel in den Stationsablauf integriert worden. Die Studierenden wussten, dass zwei Minuten vor Ablauf der Zeit ein Zeichen gegeben wurde, um ihnen die Möglichkeit zu geben, ein adäquates Gesprächsende zu finden. Die geringeren Bewertungen im Bereich Gesprächsende können jedoch auch ein Hinweis sein, dass Aspekte des Gesprächsendes im bisherigen Studium nicht ausreichend vermittelt wurden. Im Kommunikationstraining sollte auf den Punkt „Klärung noch offener Fragen“ deutlicher hingewiesen werden. Im Rahmen einer im RSM durchgeführten Studie zur Validierung eines Global Rating zur Prüfung kommunikativer Kompetenzen ^[82] wurden Studierenden zusätzlich mit einer kurzen Version des Cambridge-Calgary-Observation-Guide ^[8] bewertet. Als nicht veröffentlichtes Nebenergebnis dieser Studie zeigte sich ebenfalls, dass Items bezüglich des Gesprächsabschlusses, dabei vor allem die Klärung offener Fragen sowie der Gesprächsabschluss durch eine Zusammenfassung, selten vollständig erfüllt wurden. Es wurde die Schlussfolgerung gezogen, auf Aspekte des Gesprächsendes im Training deutlicher hinzuweisen.

4.2 Diskussion der Fragestellung

Zur Klärung der Frage, inwieweit sich die Studierenden der drei Jahrgänge in ihren kommunikativen Fähigkeiten unterscheiden, wurde eine Varianzanalyse der Mittelwerte der Items der Prüferbewertung gerechnet. (s. Kap. 3.4) Die Varianzanalyse zeigt einen signifikanten Haupteffekt für den Faktor Semester. Damit ist ein signifikanter Unterschied in Abhängigkeit vom aktuellen Semester bei der Bewertung der Studierenden mit der Prüfer-Checkliste nachgewiesen. Der Post-Hoc-Vergleich (Tab. 8) zeigt jedoch, dass die Bewertung der einzelnen Semester nicht in der erwarteten Reihenfolge erfolgte. Die Studierenden des dritten Semesters erhielten zwar signifikant höhere Bewertungen von den Prüferinnen als die des ersten Semesters, jedoch auch signifikant höhere Bewertungen als die des fünften Semesters.

Verschiedene Erklärungen für diese nicht erwartete Reihenfolge sind denkbar. Kraan et al. weisen bei einer ähnlichen Studie auf drei Voraussetzungen hin, die nötig sind, um aus quasi-

experimentellen Querschnittsdaten heraus longitudinale Aussagen zu folgern.^[12] Erstens muss auf Selektionseffekte bei den untersuchten Kohorten geachtet werden. Zweitens ist auf störende Effekte zu achten, die vom Instrument verursacht sind, und drittens sollte das „treatment“ in allen Studienjahren konstant sein. Veränderungen im Unterricht und unterschiedliche Schwerpunktsetzungen in den Studienjahren können das Ergebnisse beeinflussen.

Analog dazu können folgende Fragen formuliert werden:

- Gab es Selektionseffekte bei den drei untersuchten Kohorten? (s. Kap. 4.2.1)
- Welche mögliche Störfaktoren führten zur Beeinflussung der Ergebnisse? (s. Kap. 4.2.2)
- Gab es Veränderungen im Konzept der *Übung Interaktion* zwischen dem ersten Jahrgang (hier fünftes Semester) und dem zweiten Jahrgang (hier drittes Semester), die Ursache für die unerwartete Reihenfolge der Bewertung sein können? (s. Kap. 4.2.3)

Welche sonstige Interpretation das Ergebnis der Varianzanalyse zulässt wird anschließend im Kapitel 4.2.4 diskutiert.

4.2.1 Selektionseffekte

Da die Studienanfänger der Jahre 1999, 2000 und 2001 zum Großteil von der ZVS (Zentralstelle für die Vergabe von Studienplätzen) der Charité zugeteilt wurden und die Studierenden des Reformstudienganges jedes Jahr per Los ermittelt und nicht gezielt ausgesucht wurden, können vergleichbare Gruppen vermutet werden. Dennoch bestehende Selektionseffekte sind bei den drei untersuchten Kohorten, den Studierenden der Semester 1, 3 und 5, am wahrscheinlichsten anhand Geschlechtsverteilung und Altersstruktur der Jahrgänge festzustellen. Tab. 1 präsentiert eine ähnliche Geschlechtsverteilung in allen drei Jahrgängen mit jeweils mehr Frauen als Männern. Der Männeranteil des ersten Semesters ist bei den in die Auswertung eingehenden Studierenden etwas größer als in den anderen Semestern (+10%), da 9 der 11 nicht teilnehmenden bzw. das Gespräch abbrechenden Studierenden Frauen waren. Die Varianzanalyse bezüglich der Mittelwerte der Items der Prüfer- und der SP-Bewertung konnte weder einen Haupteffekt für den Faktor Geschlecht noch eine Interaktion mit dem Faktor Semester nachweisen und schließt somit eine Störung durch die Variable Geschlecht weitgehend aus. Einflüsse des Geschlechts der Studierenden auf kommunikative Fähigkeiten werden im Kapitel 4.4 diskutiert. Bei Betrachtung der Altersstruktur in den Jahrgängen fällt auf, dass der Altersmedian der Studierenden in höheren Semestern natürlicherweise größer ist als in unteren Semestern. Das Maximum des ersten und dritten Semesters ist durch wenige „ältere“ Ausreißer höher. Der Großteil der Studierenden aller drei Kohorten ist zwischen 20 und 25 Jahre alt. In der

Alterstruktur unterscheiden sich die drei Jahrgänge folglich nicht wesentlich. Selektionseffekte durch die Altersverteilung der Studierenden können bei ähnlicher Altersstruktur der drei Kohorten weitgehend ausgeschlossen werden.

4.2.2 Störfaktoren

Störende Effekte des Messinstruments sind bei identischer Aufgabenstellung, Stationsaufbau und OSCE-Integration am ehesten durch Unterschiede bei Simulationspatientinnen bzw. Prüferinnen zu erwarten. Die Verteilung der SP erfolgte gleichmäßig auf alle Semester. (s. Tab. 3) Das ausführliche und sorgfältige Training der SP sollte eventuelle Unterschiede in der Rollendarstellung, die zu abweichenden Bewertungen der Studierenden führen könnte, vorbeugen. Eventuelle Unterschiede bei der Rollendarstellung werden jedoch bei gleichmäßiger Verteilung der SP auf alle Semester gleichmäßig verteilt. Eine Interaktion zwischen den Faktoren SP und Semester kann somit weitgehend ausgeschlossen werden.

Da bei einer Voruntersuchung im Rahmen der Schulung der Prüferinnen mit der Prüfer-Checkliste eine Interrater-Reliabilität von $r = 0,75$ erzielt werden konnte,^[81] wurde von einer genügenden hohen Konstanz der Prüferbewertung in der Kommunikationsstation ausgegangen. Die Interrater-Reliabilität konnte im OSCE nicht verifiziert werden, da die Bewertung der Studierenden in den OSCE-Stationen nur von einer einzigen Prüferin erfolgte. Aus organisatorischen und terminlichen Gründen konnten die Prüferinnen nicht ebenso gleichmäßig wie die SP auf alle Semester verteilt werden. (s. Tab. 4) Eine Beeinflussung der Ergebnisse durch die inhomogene Verteilung der Prüferinnen kann nicht vollkommen ausgeschlossen werden. Ein Hinweis für Abweichungen in der Prüferbewertung liefert die geringere Korrelation zwischen SP- und Prüferbewertung im fünften Semester im Gegensatz zum dritten und ersten Semester. (s. Tab. 9) Solche Divergenzen können eine Ursache für die unerwartete Reihenfolge zwischen den Semesterergebnissen sein. Um bei zukünftigen Untersuchungen diese Fehlerquelle zu beseitigen, sollte eine gleichmäßige Verteilung der Prüfer/-innen auf alle Semester erfolgen bzw. eine Bewertung der Studierenden von mehreren unabhängigen Prüferinnen/-n an mehreren Stationen ermöglicht werden, um Bewertungsunterschiede weitestgehend auszugleichen.

Da der OSCE für jedes Semester an jeweils einem Tag stattfand, konnte nicht verhindert werden, dass sowohl SP wie auch Prüferinnen in der Kommunikationsstation über das aktuelle Semester der Prüflinge informiert waren. Eine Beeinflussung der Bewertung durch das Wissen über das aktuelle Semester der Studierenden kann nicht ausgeschlossen werden. Bezüglich des Ergebnisverlustes zwischen drittem und fünften Semester hätte dieser Effekt jedoch eher eine entgegengesetzte Wirkung gehabt.

4.2.3 Veränderungen im Lehrkonzept

Die Erfahrungen und die Kritik, die in den ersten Semestern der Übung Interaktion gesammelt wurden, führten in der Anfangsphase des RSM zu regelmäßiger Überarbeitung und Optimierung des Lehrkonzeptes der Übung Interaktion. Nach dem ersten Jahr des ersten Jahrganges (hier fünftes Semester) wurde von Studierenden der Wunsch nach „Konkretisierung des Konzeptes und der angebotenen Themen“ bzw. nach einer stärkeren „Vermittlung von theoretischen Aspekten zu Kommunikation und mehr praktische Übungen“ geäußert.^[83] Zum ersten Semester des zweiten Jahrganges (hier drittes Semester) erfolgte eine Überarbeitung des Konzeptes mit einer Straffung und deutlicheren zeitlichen Strukturierung der Inhalte der Übung Interaktion, um den Studierenden eine bessere Orientierung zu ermöglichen. Im Jahr 2001 wurde schließlich die Arbeitsgruppe Interaktionskonzepte (AGIK) gegründet, um ein übergeordnetes Gesamtkonzept mit aufeinander aufbauenden Semesterkonzepten zu entwickeln, welches ab dem Wintersemester 2002/2003 in die Praxis umgesetzt wurde.

Die fehlende Erfahrung zu Beginn des RSM kann zu Defiziten in der Anfangsphase der Übung Interaktion geführt haben. Die Studierenden des zweiten Jahrganges (hier drittes Semester) haben eventuell von der Überarbeitung des Interaktionskonzeptes nach dem ersten Jahr des RSM derartig profitiert, dass sie bezüglich der kommunikativen Fähigkeiten in der Kommunikationsstation bessere Bewertungen erzielen konnten als die ein Jahr weiter fortgeschrittenen Studierenden des fünften Semesters.

Eine andere Interpretation lässt die Betrachtung der Themenschwerpunkte der Übung Interaktion des Wintersemester 2001/2002 zu. Die Studierenden des ersten Semesters beschäftigten sich vor allem mit Grundlagen der Kommunikation, Teamarbeit und Feedback, das dritte Semester hatte die Schwerpunkte Einführung in die Gesprächsführung und Sexualanamnese während die Studierenden des fünften Semesters sich primär mit psychosozialen Aspekten im ärztlichen Gespräch und Konzepten der Psychosomatik befassten. Da sich die Studierenden des dritten Semesters primär mit den Grundlagen der ärztlichen Gesprächsführung auseinandersetzten, waren ihnen diese zum Zeitpunkt der Untersuchung präsenter als den Studierenden des fünften Semesters, was ebenfalls den Vorsprung des dritten Semesters gegenüber dem fünften Semester in der Prüferbewertung der Kommunikationsstation erklären könnte.

4.2.4 Interpretation des Ergebnisses der Varianzanalyse

Wie lassen sich die Unterschiede zwischen den Semestern interpretieren, wenn Störeffekte und Einflüsse durch Veränderungen im Unterricht vernachlässigt werden? Nach drei Semestern erreicht die Bewertung der Studierenden mit der Prüfer-Checkliste einen Höhepunkt. Die

Studierenden werden im Mittel deutlich über „akzeptabel“ (3,7) beurteilt. Danach nimmt die Bewertung der Studierenden leicht ab. Der deutliche Anstieg des Mittelwerts der Ergebnisse der Prüferbewertung vom ersten zum dritten Semester lässt eine Vermittlung der wesentlichen Grundkenntnisse der Arzt-Patienten-Kommunikation zum dritten Semester hin vermuten. Dieser positive Trend setzt sich vom dritten zum fünften Semester hin nicht fort. Die leichte Abnahme zwischen drittem und fünften Semester deutet auf einen geringen Verlust der gewonnenen Fähigkeiten hin. Um jedoch Aussagen darüber zu treffen, ob sich dieser Trend im weiteren Verlauf fortsetzt oder die Ergebnisse wieder ansteigen, sind Überprüfungen zu späteren Zeitpunkten nötig. Eine Studie von Kraan und Kollegen überprüfte ebenfalls in einer Querschnittstudie die Fähigkeiten der Studierenden im medizinischen Interview in den Studienjahren eins bis sechs. Es zeigte sich, dass in allen Bereichen, die auch im ersten Studienjahr geprüft wurden, vom ersten zum zweiten Studienjahr ein deutlicher Verlust mit anschließendem Anstieg der Ergebnisse stattfand. Der Verlust zwischen den Studierenden des ersten und zweiten Studienjahres wurde damit begründet, dass die Studierenden sich mit zunehmenden medizinischen Wissen auf die Lösung des medizinischen Problems konzentrieren und die Interviewführung dabei vernachlässigen. Das kontinuierliche Kommunikationstraining während des gesamten Studiums führte in den höheren Jahrgängen dagegen wieder zur Steigerung der Ergebnisse.^[12]

Um zuverlässige Aussagen über die tatsächliche Entwicklung der kommunikativen Fähigkeiten in einem Jahrgang zu machen und einen kausalen Zusammenhang mit einer Intervention (Kommunikationstraining) herzustellen, ist ein longitudinales Studiendesign mit Kontrollgruppe notwendig. Es gibt nur wenige Studien, die das Kommunikationsverhalten von Studierenden nicht nur direkt vor oder nach einer Intervention (z. B. Kommunikationstraining) sondern auch über einen längeren Zeitraum erfassen. Eine 1981 von Engler et al. publizierte Arbeit weist nach, dass kommunikative Fähigkeiten nicht nur erlernt werden, sondern auch verloren gehen. Die kommunikativen Fähigkeiten von Medizinstudentinnen/-en wurden direkt nach einem 9-Wochen-Interview-Kurs und mit einem Jahr Abstand getestet. Zwischen den beiden Messungen ergab sich eine signifikante Abnahme der Bewertung der kommunikativen Fähigkeiten.^[10] Eine Arbeit von Craig bewertete Studierende der Medizin anhand 10-minütiger Interviews vor und nach einem Interview-Kurs sowie im zweiten, dritten und vierten Folgejahr. Der Vergleich mit einer Kontrollgruppe zeigte eine signifikante Verbesserung der Teilnehmer/-innen des Interviewkurses zwischen den Messungen vor und nach der Intervention. In den folgenden drei Jahren verschlechterten sich beide Gruppen. Die letzte Messung im vierten Jahr konnte keinen signifikanten Unterschied mehr zwischen Versuchs- und Kontrollgruppe feststellen.^[11] Auch

Bögels zeigt in einer 1996 veröffentlichten Studie, dass die Bewertungen sechs Monate nach einem Kurs für Interviews mit psychiatrischen Patientinnen/-en signifikant schlechter wurden als direkt nach dem Kurs.^[84] Eine Untersuchung von Pfeifer beschreibt ebenfalls einen anfänglichen Anstieg der Fähigkeiten zur Interviewführung während eines Kurses zu Beginn des Studiums. Bei einer Kontrolle nach vier Jahren kam es jedoch wieder zu einem Verlust dieser Fähigkeiten.^[85] Pfeiffer begründete dies mit der Abkehr der Studierenden von den kommunikativen Fertigkeiten hin zu den „harten wissenschaftlichen Fakten“ während des klinischen Abschnitts des Studiums bzw. mit der „culture of medicine“, der die Studierenden während dieser Zeit ausgesetzt sind. Ein anderes Ergebnis lieferte eine Untersuchung von Maguire et al. aus England. Der Erfolg von Feedback-Training wurde im Vergleich zu konventionellem Interviewtraining während des Psychiatriepraktikums untersucht. Die Studierenden führten direkt vor und nach dem Training jeweils ein Interview und fünf Jahre später ein weiteres Interview durch. Die Feedback-Gruppe hatte einen deutlich steileren Zuwachs in der Bewertung direkt nach dem Training. Über die folgenden fünf Jahre stiegen jedoch die Ergebnisse beider Gruppen an.^[86] Bögels vermutete den Grund für das von seiner Arbeit abweichende Ergebnis darin, dass das Training in Maguires Publikation nicht zu Beginn des Studiums, sondern während des klinischen Abschnitts stattfand und somit die Bedeutung des Interview-Trainings eher im Bewusstsein blieb.^[84]

Die Ergebnisse der zitierten Untersuchungen legen nahe, dass eine kontinuierliche Vermittlung kommunikativer Fähigkeiten im Medizinstudium Vorteile gegenüber punktuellen Interventionen besitzt. Der Stellenwert der Arzt-Patienten-Interaktion bleibt den Studierenden während des Studiums präsent. Die in dieser Arbeit beobachtete leichte Abnahme der Ergebnisse vom dritten zum fünften Semester bekräftigen dies. Bei einer nicht regelmäßig stattfindenden Übung Interaktion wäre der Verlust eventuell stärker ausgefallen. Eine regelmäßige Erfassung kommunikativer Fähigkeiten zeigt die tatsächliche Entwicklung dieser Kompetenzen und gibt den Studierenden und Lehrenden das nötige Feedback, um das eigene Lernen bzw. Lehrkonzepte zu überdenken.

4.3 Bewertung durch die Simulationspatientinnen

Der Vergleich der Boxplots der drei Jahrgänge der SP- (s. Abb. 9) und der Prüferbewertung (s. Abb. 8) lässt bei der Prüferbewertung eine deutlichere Abstufung der Jahrgänge erkennen. Die Mediane der SP-Bewertung steigen vom ersten zum dritten Semester kaum an. Es kommt jedoch zur deutlichen Abnahme der Spannweite. Dies deutet darauf hin, dass die „schlimmsten“ kommunikativen Fehler gegenüber den Simulationspatienten vom ersten Semester zum dritten

Semester beseitigt werden konnten.

Die Varianzanalyse des SP-Global Rating weist im Gegensatz zur Prüfer-Checkliste keinen Haupteffekt für den Faktor Semester nach. Das SP-Global Rating kann keinen signifikanten Unterschied in der Bewertung der Studierenden nach aktuellem Semester feststellen. Wie der Vergleich der Boxplots andeutet, differenziert die SP-Bewertung schlechter in Bezug auf das Semester der Studierenden als die Prüferbewertung. Die Prüfer-Checkliste ermöglicht gegenüber dem SP-Global Rating in dieser Kommunikationsstation eine deutlichere Differenzierung und kann mit ihrer Detaillierung auf spezifische Mängel (Gesprächsende) der Studierenden aufmerksam machen.

Zur Validierung der Prüferbewertung wurde eine Korrelation mit der SP-Bewertung ermittelt. (s. Tab. 9) Über alle Semester ist eine signifikante Korrelation nachweisbar, die in den Randbereichen der Prüferbewertung deutlich ansteigt. Da Prüferinnen und SP in der Bewertung von besonders „schlechten“ und „guten“ Studierenden eher übereinstimmen als bei Studierenden im mittleren Bereich, sind in den Randbereichen höhere Korrelationen zu erwarten. Dies ist auch ein Grund für höhere Korrelation zwischen Prüfer- und SP-Bewertung im ersten Semester, da die Ergebnisse dort eine größere Spannweite haben und Prüferinnen bzw. SP die Leistung der Studierenden besser differenzieren können und sich schlechtere von besseren Leistungen deutlicher abgrenzen lassen. Die Korrelation wird vom dritten zum fünften Semester noch geringer, was als Hinweis auf Unregelmäßigkeiten bei der Prüferbewertung im dritten bzw. fünften Semester gewertet werden kann. (s. Kap. 4.2.2)

Finlay et al. führten ebenfalls eine Untersuchung durch, in der kommunikative Fähigkeiten in einem simulierten Arzt-Patienten-Gespräch sowohl von erfahrenen Prüferinnen/-n als auch von SP erfasst wurden.^[49] Mit jeweils ähnlichen Bewertungsbögen wurden über drei Jahre hinweg Korrelationskoeffizienten zwischen $r = 0,47$ und $r = 0,60$ erzielt, wobei die SP bei über 80% der Kandidaten höhere Punktzahlen vergaben. Ihre eigene Bewertungsleistung betrachteten die SP teilweise kritischer als die Kandidatenleistung. Die Studie kommt zum Ergebnis, dass die SP-Bewertung für die Validierung der Prüferbewertung nützlich war, jedoch nicht die Prüferbewertung ersetzen sollte. Eine 1999 publizierte Arbeit von Regehr und Kollegen verglich ebenfalls die Bewertung von Prüferinnen/-n und SP, sowohl anhand einer Checkliste wie auch eines Global Rating.^[50] Die Kandidaten (33 *clinical clerks* im dritten Jahr und 15 *residents*) wurden an fünf Stationen geprüft. Der Korrelationskoeffizient zwischen SP- und Prüferbewertung betrug bei allen Stationen mit der Checkliste im Mittel 0,49 und mit dem Global Rating 0,43. Das Gesamtergebnis über alle fünf Stationen korrelierte allerdings bei der Checklisten mit $r = 0,70$ und beim Global Rating mit $r = 0,71$. Abweichungen in der Bewertung

an einzelnen Stationen relativierten sich über mehrere Stationen hinweg. Bei dieser Studie konnten sowohl Prüfer/-innen als auch SP signifikant zwischen *clerks* und *residents* unterscheiden, allerdings nur anhand des Global Rating. Eine 2004 veröffentlichte Arbeit untersuchte ebenfalls mehrfach Korrelationen zwischen SP-Bewertung und Experten-Bewertung, sowohl anhand von Checkliste wie auch von Global Rating.^[51] Die geringste Korrelation ergab sich zwischen Experten-Checkliste und SP-Global Rating mit $r = 0,52$. Die Korrelation zwischen Experten-Checkliste und SP-Checkliste bzw. Experten-Global Rating und SP-Global Rating war mit $r = 0,58$ bzw. $r = 0,69$ etwas größer. In der Studie wurde außerdem ein starker unabhängiger Effekt des Fall-Szenarios auf die individuelle Leistung der Studierenden deutlich.

In den zitierten Arbeiten werden SP-Bewertungen einerseits zur Validierung von Prüferbewertungen herangezogen und andererseits als deren Ersatz diskutiert. Die Korrelationen dieser Arbeit sind vergleichbar mit denen der zitierten Arbeiten. Des weitern liefern sie Hinweise für Abweichungen in der Bewertung des dritten bzw. fünften Semesters. Da die SP-Bewertung im Gegensatz zur Prüferbewertung nicht zwischen den Semestern differenzieren kann, kann sie in dieser Arbeit die Prüferbewertung nicht ersetzen.

4.4 Einflüsse des Geschlechts von Studierenden, Prüfer/in und SP

Einige Untersuchungen konnten einen Einfluss des Geschlechts der Studierenden auf gemessene kommunikative Fähigkeiten nachweisen. (s. Kap. 1.4.3) Aspegren beschreibt in einem Review, dass Frauen nach einem Kommunikationstraining bessere Ergebnisse erzielen als Männer.^[3] Roter und Kollegen veröffentlichten 2004 eine Arbeit, die den Einfluss des Studierendengeschlechts auf Veränderungen in der Kommunikation nach einem Kommunikationstraining untersuchte. Sie kamen ebenfalls zu dem Ergebnis, dass Frauen größere positive Veränderungen erlangen als Männer.^[87] Eine im selben Jahr veröffentlichte Arbeit von Wiskin et al. konnte bei der Prüfung kommunikativer Fähigkeiten einen signifikanten Zusammenhang zwischen Geschlecht und Performance in einigen aber nicht in allen Stationen feststellen, wobei Frauen besser abschnitten. Obwohl dies nicht alle Stationen betraf, hatte der Faktor Geschlecht dennoch einen Effekt auf das Gesamtergebnis.^[88]

Die Varianzanalyse der Mittelwerte der Items der Kommunikationsstation zeigt weder bei der SP- noch bei der Prüferbewertung ein Haupteffekt des Faktors Geschlecht der Studierenden bzw. eine Interaktion des Faktors Geschlecht mit dem Faktor Semester. Die Mittelwerte der Items von Studentinnen und Studenten unterscheiden sich kaum. (s. Tab. 5) Es offenbaren sich keine spezifischen Items, bei denen deutliche Unterschiede zwischen den Mittelwerten von Frauen und Männer vorhanden sind. Eine Untersuchung von Colliver über Einflüsse des Geschlechts auf die

Bewertung klinischer Kompetenz konnte ebenfalls keinen generellen Unterschied in der Bewertung von Frauen und Männern feststellen. Lediglich bezüglich der persönlichen Art und Weise (personal manner) bekamen Frauen signifikante, geringfügig bessere Bewertungen.^[89] Da im RSM im Gegensatz zu Untersuchungen wie der von Roter^[87] nicht ein einmaliges Kommunikationstraining stattfindet, sondern kommunikative Fähigkeiten über das gesamte Studium regelmäßig und stetig in der Übung Interaktion vermittelt werden, bilden sich geschlechtsspezifische Unterschiede eventuell nicht so deutlich aus, dass sie vom Prüfungsinstrument erfasst werden können.

Inwieweit sich das Geschlecht der Prüfern/innen bzw. der SP auf die Bewertung der Studierenden auswirkt, konnte nicht untersucht werden, da in der Kommunikationsstation nur Prüferinnen und Simulationspatientinnen eingesetzt wurden. Einige Publikationen untersuchten die Interaktion zwischen SP-Geschlecht und Studierendengeschlecht bei der Bewertung von Studierenden durch SP. Zwei Arbeiten von Furman^[72] und von Chambers^[73] konnten einen Haupteffekt für das Geschlecht der Studierenden und einen Haupteffekt für das SP-Geschlecht nachweisen mit jeweils besseren Bewertungen von weiblichen Studierenden bzw. besseren Bewertungen durch weibliche SP. Zwei weitere Arbeiten von Colliver^[89] und Gispert^[90] konnten nur einen Haupteffekt für das SP-Geschlechts nachweisen. Bei allen vier Arbeiten gab es jedoch keine Interaktion zwischen den Faktoren Studierenden- und SP-Geschlecht. Vor diesem Hintergrund kann vermutet werden, dass die Ergebnisse der Kommunikationsstation qualitativ kaum dadurch beeinflusst wurden, dass lediglich Prüferinnen bzw. Simulationspatientinnen zum Einsatz kamen.

4.5 Reliabilität und Validität

Der Reliabilitätskoeffizient Cronbach's Alpha stellt die innere Konsistenz einer additiven Skala dar und ist somit ein Wert für die Homogenität eines Bewertungsbogens. Ein Cronbach's Alpha von $\alpha = 0,88$ bei der Prüfer-Checkliste und $\alpha = 0,84$ beim SP-Global Rating weisen für beide Messinstrumente eine ausreichende innere Konsistenz nach. Der Zielbereich für Cronbach's Alpha liegt zwischen $\alpha = 0,70$ und $\alpha = 0,90$.^[91] Niedrigere Werte deuten auf zunehmende Inhomogenität im Bewertungsbogen hin während höherer Werte bedeuten können, dass unterschiedliche Items dieselbe Frage in nur leicht veränderter Art und Weise darstellen.

Im Vorfeld der Studie wurde eine Interrater-Reliabilität von $r = 0,75$ bestimmt, die für den Zweck dieser Studie ausreichend ist. Da die Studierenden jeweils nur von einer Prüferin bewertet wurden, konnte dieser Wert im OSCE nicht überprüft werden. Die nicht regelmäßige Verteilung der Prüferinnen auf alle Semester im Zusammenhang mit der nicht erwartete Reihenfolge der

Ergebnisse zwischen drittem und fünftem Semester und die sinkende Korrelation zwischen SP-Global Rating und Prüfer-Checkliste zwischen drittem und fünftem Semester deuten jedoch auf Unterschiede bei der Bewertung durch die Prüferinnen hin. Bei einer weiteren Verwendung der Messinstrumente sollte diese Problematik beachtet werden. Außerdem empfiehlt sich das Prüfen über mehrere Stationen und somit von mehreren Prüferinnen/-n, um Abweichungen bei den Prüferinnen/-n im Gesamtergebnis zu minimieren.

Zur Wahrung der Inhaltsvalidität erfolgte die Prüfung an einem realistischen Fall mit einer dafür entwickelten Checkliste. Die Checkliste bzw. das Global Rating wurde mit Hilfe von bewährten in der Literatur beschriebenen Bewertungsbögen entwickelt. (s. Kap. 2.3.1) Die signifikant positive Korrelation zwischen SP- und Prüferbewertung von $r = 0,38$ dient als Hinweis für Kriteriumsvalidität und liegt im Bereich von vergleichbaren publizierten Arbeiten. (s. Kap. 4.3) Letztendlich konnte das Messinstrument signifikant zwischen den Semestern differenzieren, allerdings nur in der erwarteten Reihenfolge zwischen erstem und drittem bzw. zwischen erstem und fünftem Semester.

4.6 Aussagekraft des Evaluationsverfahren

Die Messinstrumente verfügen über eine gute innere Konsistenz. Die Prüferbewertung konnte zwischen den Semestern differenzieren. Es bestehen jedoch Hinweise auf Abweichungen innerhalb der Prüferbewertung beim dritten und fünften Semester. Die Aussagekraft dieser Studie über die kommunikativen Fähigkeiten der Studierenden ist jedoch vor allem begrenzt, da diese aus organisatorischen Gründen nur an einer Station geprüft wurden. Das Ergebnis an einer Station kann nicht zwingend als Vorhersage für die kommunikativen Fähigkeiten der Studierenden an anderen Stationen bzw. in anderen Situationen verwendet werden. Mehrere Publikationen zeigen, dass mit einer einzigen Station kommunikative Fähigkeiten nicht exakt erfasst werden können.^[66, 67, 92] Das gemessene Phänomen ist abhängig vom Fall und von der Aufgabenstellung. Es werden Prüfungszeiten von 2 bis 3 Stunden in mehreren Stationen benötigt werden, um den Einfluss der Fallspezifität ausreichend zu verringern und genügend hohe Reliabilitäten zu erreichen. (s. Kap. 1.4.2) Bei genügend hohen Fallzahlen gleichen sich Unterschiede der Bewertung bei unterschiedlichen Stationen bzw. Prüfern aus.^[27] Dies wird beispielsweise an der im Zusammenhang mit der SP-Bewertung zitierten Arbeit von Regehr und Kollegen deutlich. Die Korrelation zwischen SP- und Prüferbewertung betrug im Mittel an den Stationen $r = 0,49$. Die Gesamtergebnisse über fünf Stationen korrelierten dann immerhin mit $r = 0,70$.^[50] Aus zeitlichen, strukturellen und finanziellen Gründen war im Rahmen der Untersuchung der kommunikativen Fähigkeiten der Studierenden des RSM im Wintersemester

2001/2002 die Integration mehrerer „Kommunikationsstationen“ in die OSCEs der Semesterabschlussprüfungen nicht möglich. Für weitere Untersuchungen kommunikativer Fähigkeiten am RSM wäre die Durchführung eines speziellen Kommunikations-OSCE sinnvoll. Dies ist jedoch mit zusätzlichem zeitlichen, personellen und finanziellen Aufwand verbunden, der zurzeit kaum realisierbar ist.

Diese Studie weist darauf hin, dass zu Beginn des RSM deutliche Fortschritte in Bezug auf die kommunikativen Fähigkeiten der Studierenden gemacht wurden. Zwischen dem ersten Semester, das als „Kontrollgruppe“ mit geringer Intervention interpretiert werden kann, und dem dritten bzw. fünften Semester konnten deutliche Unterschiede erfasst werden. In höheren Semestern waren die Unterschiede geringer. Die Prüfer-Checkliste kann in den höheren Ergebnisbereichen Unterschiede in kommunikativen Kompetenzen nicht mehr deutlich genug abbilden. Vor weiteren Untersuchungen zur Erfassung kommunikativer Fähigkeiten sollte eine genaue Evaluierung und Überarbeitung der Messinstrumente mit Hilfe der bei dieser Studie gewonnenen Daten erfolgen, um eine kontinuierliche Erfassung der kommunikativen Fähigkeiten zu ermöglichen und eine reliable und valide Differenzierung sicherzustellen. Um jedoch einen klaren kausalen Zusammenhang zwischen einer bestimmten Intervention, z. B. Übung Interaktion, und einem Zuwachs an kommunikativen Fähigkeiten herzustellen, wird eine echte Kontrollgruppe benötigt, die in der Praxis jedoch schwer zu formieren ist.

4.7 Ausblick

Im Rahmen dieser Arbeit wurde im Wintersemester 2001/2002 im Reformstudiengang Medizin der Charité erstmals versucht, kommunikative Fähigkeiten der Studierenden aller drei Jahrgänge zu erfassen. Während zwischen erstem und drittem Semester ein deutlicher Anstieg der kommunikativen Fähigkeiten erkennbar war, ist die weitere Entwicklung nach dem dritten Semester nicht deutlich geworden. Für die Dokumentation der weiteren Entwicklung kommunikativer Fähigkeiten während des Studiums und für die Weiterentwicklung des Curriculums im Hinblick auf die Vermittlung dieser Fähigkeiten ist eine regelmäßige Erfassung der kommunikativen Fähigkeiten der Studierenden notwendig.

Bei der Auswertung der Ergebnisse dieser Arbeit wurden die Grenzen dieser Untersuchung deutlich. Eine gründliche Evaluierung und Optimierung der Methode und der Messinstrumente ist vor einer weiteren Anwendung notwendig. Die bei dieser Studie aufgenommenen Videoaufnahmen können dabei eine Hilfe bei der gründlichen Evaluierung der Methode sein. Wesentliche Schwierigkeiten dieser Studie sind die ungleiche Verteilung der Prüferinnen auf die Semester und die Prüfung an einer einzigen Station. Der zukünftige Einsatz einer einzigen

zusätzlichen OSCE-Station ist aus Gründen der Reliabilität und Fallspezifität nicht sinnvoll. Ein eigener „Kommunikations-OSCE“ zu mehreren Zeitpunkten im Studium, in dem die Studierenden an mehreren Stationen von verschiedenen Prüferinnen/-n bezüglich ihrer kommunikativen Fähigkeiten bewertet werden, wäre eine optimale Lösung. Wegen des notwendigen finanziellen und personellen Aufwandes ist dies zurzeit jedoch nicht realisierbar.

Als günstigere Alternative bietet es sich an, die Checklisten von bestehenden OSCE-Stationen der Semesterabschlussprüfungen mit Global Ratings zu ergänzen, die sich auf die kommunikativen Fähigkeiten der Studierenden beziehen. Dies wurde im Reformstudiengang mittlerweile schon mehrfach probiert und lieferte zum Teil vielversprechende Ergebnisse. Im Wintersemester 2003/2004 konnte eine weitere Studie im RSM zeigen, dass zusätzliche Global Ratings für die Erfassung kommunikativer Kompetenzen in vorhandenen Anamnesestationen eines bestehensrelevanten OSCE reliable und valide Ergebnisse erzielen können.^[82] Eine wichtige Voraussetzung ist jedoch eine gute Schulung der Prüfer/-innen und die Prüfung über mehrere Stationen. Als Erfolg der Bemühungen, die Überprüfung kommunikativer Fähigkeiten im RSM voranzutreiben, kann der Beschluss des Prüfungsausschusses des RSM gewertet werden, dass bei den Semesterabschlussprüfungen des achten Semesters im Sommersemester 2005 die Ergebnisse zusätzlicher Global Ratings für die Prüfung kommunikativer Fähigkeiten zu 30% in das Ergebnis der jeweiligen OSCE-Station eingehen.^[93]

Der Nutzen für das Prüfungsprogramm als Ganzes darf bei der Implementierung von Prüfungen zur Erfassung kommunikativer Fähigkeiten jedoch nicht außer Acht gelassen werden. Van der Vleuten und Schuwirth vertreten in einem jüngst erschienen Artikel zur Prüfung professioneller Kompetenzen im Medizinstudium die Ansicht, dass die Prüfungsliteratur sich überwiegend an individuellen Prüfungsmethoden orientiert und sich zu sehr auf exklusive psychometrische Eigenschaften wie z. B. Reliabilität und Validität konzentriert. Es gibt keine grundsätzlich schlechten oder guten Prüfungsmethoden, jede Methode kann nützlich sein in Abhängigkeit ihrer Anwendung und des programmatischen Kontextes. Unsere Aufmerksamkeit und Bemühen sollte hauptsächlich darauf fokussieren, das Prüfungsprogramm als integrierten Teil des Curriculums zu sehen.^[27] Es bleibt abzuwarten, inwieweit die Beschlüsse des Prüfungsausschusses des RSM Einfluss auf die Bedeutung der Übung Interaktion und das Lernverhalten der Studierenden nehmen.