Aus der Abteilung Anaesthesiologie I

(Komm. Leiter: Prof. Dr. med. M. Quintel)

im Zentrum Anaesthesiologie, Rettungs- und Intensivmedizin

der Medizinischen Fakultät der Universität Göttingen

Entwicklung, Reliabilität und Objektivität einer „Objective Structured Clinical Examination“ in der Notfallmedizin

INAUGURAL - DISSERTATION

zur Erlangung des Doktorgrades

der Medizinischen Fakultät

der Georg-August-Universität zu Göttingen

vorgelegt von

Katrin Schwerdtfeger

aus

Seesen

Göttingen 2010

D e k a n:	**Prof. Dr. med. C. Frömmel**
I. Berichterstatter:	**Priv. Doz. Dr. med. A. Timmermann**
II. Berichterstatter/in:	
III. Berichterstatter/in:	
Tag der mündlichen Prüfung:	

I. INHALTSVERZEICHNIS

II. Abkürzungen, Abbildungen und Tabellen

II.1 Liste der verwendeten Abkürzungen

Abb.	Abbildung
ÄAppO	Approbationsordnung für Ärzte
Airway	Station „Sicherung des Atemweges“
BLS	Station Basic- Life- Support
bzw.	beziehungsweise
Defi	Station Defibrillation
DV	Digital Video
ERC	European Resuscitation Council
evtl.	eventuell
F	Item der Checkliste
κ	Korrelationskoeffizient Kappa nach Cohen
MCQ	Multiple Choice Questions
n	Anzahl
OSCE	Objective Structured Clinical Examination
Pkt.	Punkte
Rhythmus	Station Rhythmusdiagnostik
r_s	Spearmans Rangkorrelationskoeffizient
SAS	Statistic Analysis Software (Statistikprogramm)
SD	Standardabweichung
SPSS	Statistical Package for the Social Sciences (Statistikprogramm)
Stud	Videoratergruppe Studenten
Tab.	Tabelle
Trauma	Station “Versorgung eines verunfallten Patienten“
VHS	Video Home System
z.B.	zum Beispiel

II.2 Abbildungsverzeichnis

II.3 Tabellenverzeichnis

1. EINLEITUNG

1.1 Ausgangssituation

Mit Einführung der neuen Approbationsordnung für Ärzte (ÄAppO) (Bundesministerium für Gesundheit 2002; Georg-August-Universität Göttingen 2004) erfolgten ab dem Sommersemester 2004 durch die Universitäten zahlreiche Umstrukturierungen und Neuerungen. Hinsichtlich der Leistungskontrollen dienten bisher fakultätsinterne Klausuren ausschließlich dem Scheinerwerb. Eine Differenzierung der Prüfungsergebnisse musste nicht erfolgen. In der neuen ÄAppO werden nun benotete Leistungsnachweise gefordert. Auf dem Abschlusszeugnis werden die Noten des klinischen Studienabschnittes gesondert ausgewiesen. Somit bekommen die Prüfungen nicht nur für die Studierenden, sondern auch für die Hochschule eine andere Qualität. Eine weitere Forderung der neuen ÄAppO ist es, dass fächerübergreifendes Denken mit Hilfe interdisziplinärer Unterrichtsformen und Unterricht in Querschnittsbereichen gefördert werden. Die praxisnahe Lehre bekommt einen höheren Stellenwert. Neben Vorlesungen sollen insbesondere praktische Übungen und Seminare in Form des Kleingruppenunterrichtes durchgeführt werden. Die Studierenden sind verpflichtet, die erfolgreiche Teilnahme an praktischen Übungen nachzuweisen. §2 ÄAppO *„Eine erfolgreiche Teilnahme an einer Übung nach Absatz 3 liegt vor, wenn die Studierenden in der praktischen Übung in einer dem betreffenden Fachgebiet angemessenen Weise gezeigt haben, dass sie sich die erforderlichen Kenntnisse, Fähigkeiten und Fertigkeiten angeeignet haben und sie in der Praxis anzuwenden wissen.“* Der Bereich Notfallmedizin ist als ein typisches interdisziplinäres Fach ein leistungsnachweisfähiger Querschnittsbereich geworden und nimmt somit einen größeren Anteil im Medizinstudium ein.

Daraus wurde als Zielvorgabe der Medizinischen Fakultät der Universität Göttingen gefordert, Prüfungsformen zu wählen, die speziell für die zu überprüfende Qualität geeignet sind (Georg-August-Universität Göttingen 2004).

Mit der Umsetzung dieser Forderung wurde im Sommersemester 2004 für den Bereich Notfall- und Intensivmedizin eine praktische Prüfung ähnlich einer Objective Stuctured Clinical Examination (OSCE) eingeführt (Timmermann et al. 2005). Während viele Autoren die Einführung von solchen praktischen Prüfungen

fordern (Beckers et al. 2004) und international OSCEs seit Jahren etabliert sind, steigt nun auch in Deutschland die Zahl der OSCE-Prüfungen in den vergangenen Jahren stetig. Neben den Universitäten Düsseldorf, Göttingen, Hannover, Münster, Ulm und Witten-Herdecke, die schon seit einiger Zeit OSCEs durchführen (Chenot und Ehrhardt 2003), sind in der Literatur auch OSCE- Prüfungen an den Universitäten in Berlin (Scheffer et al. 2008), Erlangen (Heckmann et al. 2008), Frankfurt (Ziegler und Wagner 2008), Heidelberg (Junger et al. 2005), Halle (Mau und Kusak 2005), München (Schwarzkopf et al. 2007) und Tübingen (Schrauth et al. 2006) zu finden.

Während vereinzelt über Stationen mit notfallmedizinischem Inhalt, wie Basic-Life-Support berichtet wurde (Chenot et al. 2004), gibt es bislang in Deutschland kaum Erfahrungen mit einer rein notfallmedizinisch ausgerichteten OSCE (Weißer et al. 2004). Auch international wurde die OSCE zur Prüfung des Lern- und Lehrerfolgs nur vereinzelt in der Notfallmedizin eingesetzt: Traumamanagement (Ali et al. 2002; Ali et al. 1996a; Ali et al. 1996b; Hill et al. 1997; Li et al. 2006) und allgemeine Notfallmedizin (Beckers et al. 2004; Burdick et al. 1996; Johnson und Reynard 1994; Lunenfeld et al. 1991).

In der ärztlichen Weiterbildung bestehen Erfahrungen in der Durchführung praktischer Prüfungen vor allem im Bereich der kardiopulmonalen Reanimation bei der Durchführung der Advanced Life Support und European Paediatric Life Support Kurse des European Resuscitation Councils (ERC) (Baubin und Dirks 2008).

Um dem praktischen Anspruch des Notfall- und Intensivmoduls einerseits und den Vorgaben seitens der Fakultät andererseits gerecht zu werden, soll neben einer theoretischen Prüfung, auch eine praktisch-klinische Prüfung im Sinne einer OSCE die Lernkontrolle gewährleisten. Die Studierenden sollen motiviert werden, sich praktische Fertigkeiten anzueignen, um diese in ihrer anschließenden klinischen Tätigkeit, aber auch als Ersthelfer in präklinischen Situationen anwenden zu können. Als globales Lernziel wurde die Fertigkeit definiert, eine akut lebensbedrohliche Erkrankung eines Patienten so lange versorgen zu können, bis spezialisierte Hilfe die weitere Behandlung übernehmen kann.

1.2 Prüfungen in der medizinischen Ausbildung

Im Medizinstudium, wie auch in der Fort- und Weiterbildung, bilden Prüfungen zur Evaluation des Lern- und Lehrerfolges wichtige Eckpfeiler. Sei es, um den Studierenden ihren Kenntnisstand im Prüfungsfach mitzuteilen, den Fakultäten eine Rückmeldung über die Effektivität der Ausbildung zu geben, und ebenso der Gesellschaft ein gewisses Maß an fachlicher Kompetenz zu garantieren, da bestimmtes Basiswissen vorausgesetzt wird. Das Ziel der ärztlichen Ausbildung ist laut § 1 ÄAppO (Bundesministerium für Gesundheit 2002), grundlegende Kenntnisse, Fähigkeiten und Fertigkeiten in allen Fächern zu vermitteln (Petrusa et al. 1987).

Die zertifizierenden Prüfungen für das Staatsexamen, wie Multiple-Choice-Prüfungen und unstrukturierte mündliche Prüfungen, ebenso wie Multiple-Choice-Klausuren oder mündliche Testate als Semesterprüfungen eignen sich nur bedingt, um klinische Kompetenz zu bewerten, da die Multiple-Choice-Fragen Fähigkeiten und nicht Fertigkeiten überprüfen und mündliche Prüfungen nicht reliabel genug sind (Dupras und Li 1995; Levine et al. 1970; Wass und van der Vleuten 2004). Miller stellte 1990 ein Pyramidenmodell (s. Abb. 1.1) zur Entwicklung der klinischen Kompetenz von Studierenden in Studium und Weiterbildung vor. Die Basis dieser Pyramide bildet das grundlegende theoretische Wissen (Knows), welches die Studenten benötigen. Die darüber liegenden Stufen werden vom Wissen über die praktische Umsetzung (Knows how) und der eigentlichen Anwendung (Shows how) des Gelernten gebildet. Auf der letzten Stufe des Modells sollen die Fähigkeiten und Fertigkeiten im praktischen Alltag außerhalb von Prüfungssituationen angewendet werden (Does) (Miller 1990). Ziel einer Prüfung sollte sein, auf einer möglichst hohen Stufe zu prüfen.

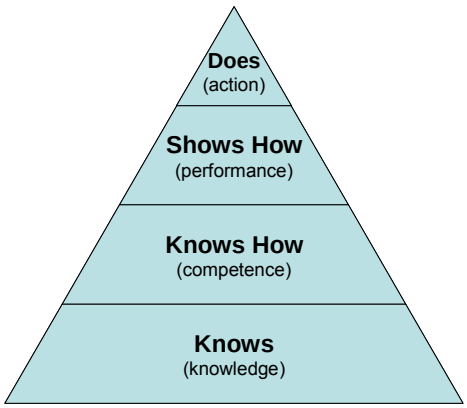

Abb.1.1 Lernpyramide nach Miller (Miller 1990; S. 63)

Eine klinisch-praktische Prüfung bietet sich als Erfolgkontrolle an, weil die Art der Prüfungstechnik einen direkten Einfluss auf die Lernstrategie besitzt (Marton und Saljo 1976) und die Leistung der Studierenden verbessert (Newble und Entwistle 1986; Petrusa et al. 1987; van der Vleuten und Schuwirth 2005). Unpassende Prüfungstechniken hingegen können in bestimmten Bereichen zu Fehleinschätzungen und Behandlungsfehlern führen, die einen Einfluss auf die Behandlungsqualität haben können (Newble 1992). Daher wird diese Art der Prüfungsform sowohl von Studierenden, als auch von Dozenten gefordert (Newble et al. 1979; Simpson 1972). Harden und Mitarbeiter führten 1975 erstmals eine solche praktische Prüfung in Form einer OSCE durch (Harden et al. 1975).

1.3 Objective Structured Clinical Examination – OSCE

Die OSCE-Prüfungsmethode wurde von Harden et al 1975 zum ersten Mal publiziert (Harden et al. 1975). OSCE steht hierbei nach Harden (Harden und Gleeson 1979; Harden et al. 1975) für **O**bjective **S**tructured **C**linical **E**xamination. Die OSCE sollte die bis dahin unstrukturierten Prüfungen am Patientenbett ersetzen, bei denen das Schwierigkeitsmaß von Prüfung zu Prüfung variiert und das Beurteilungsmaß vom Prüfer abhängig ist. Durch das Zusammenfassen unterschiedlicher Stationen und mehrerer Prüfer wird in dieser Prüfungsform die

klinische Kompetenz der Studierenden in objektiver und vor allem standardisierter Form ermittelt.

Grundstruktur einer OSCE Prüfung ist ein Parcours mit voneinander unabhängigen Stationen, durch den die Kandidaten rotieren, bis sie alle Prüfungsstationen absolviert haben. Dabei steht Ihnen für jede Aufgabe ein fester Zeitrahmen zur Verfügung, der je nach Größe der Prüfung zwischen fünf und zwanzig Minuten betragen kann. Das Spektrum der Stationen reicht dabei von Anamneseerhebung, über die klinische Untersuchung bis zur Reanimation am Phantom. Die subjektive Beeinflussung durch den Prüfer soll in der OSCE weitgehend dadurch aufgehoben werden, dass der Prüfling Kontakt mit mehren Prüfern hat und damit von einer gegenseitigen Aufhebung subjektiver Faktoren der Prüfer auf das Prüfungsergebnis auszugehen ist. Die Prüfer sind einer Station zugeteilt und bewerten die Studenten anhand einer Checkliste oder einer Globalbeurteilung z.B. in Form von Schulnoten. Die Checkliste besteht aus einer Liste detaillierter Aufgaben (Items), die von den Studierenden absolviert werden sollen. Die Kandidatenleistung wird nach dem Ja/Nein-Prinzip (erfüllt/nicht erfüllt) beurteilt. Globalbeurteilungen beinhalten im Gegensatz zur Checkliste weniger oder nur ein Item, sind allgemeiner formuliert und werden auf einer Likert-Skala mit mehreren Abstufungen bewertet (z.B. von sehr gut bis sehr schlecht auf einer Skala von 1-10). Diese Skalenbeurteilung wird vor allem für die Bewertung von kommunikativen Fähigkeiten, Verhalten und zur Beurteilung von Problemlösungsstrategien eingesetzt. McIlroy et al. zeigen, dass das Verhalten der Prüfungskandidaten sich ändert, wenn sie vorher erfahren, ob sie anhand einer Checkliste oder mittels Global Ratings bewertet werden. Prüflinge, die eine Checklistenbewertung erwarten, zeigen ein einstudiertes Ablaufschema, während Kandidaten, die eine Globalbeurteilung erwarten, mehr Wert auf Verhalten und Kommunikation legen, als auf ein standardisiertes Schema (McIlroy et al. 2002).

Um eine hohe Inhaltsvalidität sicherzustellen, sollen mit Hilfe eines Blueprints die Prüfungsinhalte mit den Zielen des zugrunde liegenden Curriculums abgeglichen werden (Newble 2004). Die Dauer und die Anzahl der Stationen sind je nach Prüfungsfach sehr unterschiedlich. In einer OSCE kann, neben Fertigkeiten, mit Hilfe von Multiple-Choice-Stationen auch Wissen abgeprüft werden. Für Gesprächsführung werden im Allgemeinen zuvor geschulte

Simulationspatienten eingesetzt (Barrows 1968; Cohen R. et al. 1990; Harden et al. 1975). Die Verbindung von Strukturiertheit und Objektivität zeigt in Studien eine hohe Reliabilität der OSCE (Cohen R. et al. 1990; Petrusa et al. 1987). Andere Autoren weisen jedoch darauf hin, dass für ein umfassendes Verständnis von klinischen Kompetenzen die OSCE-Prüfung allein nicht ausreichend ist, sondern dass andere Prüfungsmethoden zusätzlich eingebracht werden sollten (Barman 2005; Verma und Singh 1993). Newble zeigt, dass man durch die Verknüpfung von OSCE und einer schriftlichen Prüfung die Reliabilität deutlich steigern und die Prüfungsdauer verkürzen kann (Newble 2004; Newble und Swanson 1988). Durch die Kombination der Beobachtung praktischer Fertigkeiten einerseits und der schriftlichen oder mündlichen Prüfung von Wissens und Denkleistungen anderseits wird die Prüfung inhaltlich heterogener (Bloch et al. 1999).

Ein Test kann erst dann gültige Ergebnisse liefern, wenn er objektiv und reliabel ist. Diese beiden Testgütekriterien sind notwendige, aber nicht hinreichende Voraussetzungen für die Validität des Tests. Diese Qualitätsanforderungen sind nur annäherungsweise zu erreichen. Hinzu kommt, dass eine objektive, valide und reliable Prüfung sehr kostenintensiv ist. Des Weiteren ist eine OSCE-Prüfung personalintensiv. So müssen Prüfer aus evtl. unterschiedlichen Fachbereichen und Laienschauspieler als Simulationspatienten zur Verfügung stehen. Schon die Planungsphase einer OSCE-Prüfung ist weitaus aufwendiger als die Vorbereitung einer unstrukturierten mündlichen Prüfung oder eines Multiple-Choice-Tests. Neben dem Design von Prüfungsstationen und Checklisten müssen geeignete Räumlichkeiten organisiert und ausreichend Verbrauchsmaterial bereitgehalten werden. Zudem brauchen sowohl Prüfer wie auch Simulationspatienten eine ausreichende Schulung und Einweisung.

Eine OSCE, die die Qualitätskriterien zufriedenstellend erfüllt, bedingt aber nicht zwingend eine Prüfung der klinischen Kompetenz in ausreichendem Maße. So stellt die OSCE eine Prüfungssituation dar und keine „real-life“-Situation und die Studierenden können ihre Fähigkeiten nicht als Ganzes präsentieren, sondern müssen sie an Stationen stückchenweise hervorbringen (Barman 2005).

Newble und Swanson fordern darüber hinaus eine Mindestanzahl von 20 Prüfungsstationen bzw. eine mindestens vier Stunden dauernde Prüfung, bestehend aus einer OSCE und einem zusätzlichen schriftlichen Test, um die

Qualitätskriterien zu erreichen (Newble und Swanson 1988). Dies bezieht sich jedoch auf eine interdisziplinär gestaltete OSCE, die am Ende eines Studienabschnittes durchgeführt wird. In dem vorliegenden Modul handelt es sich jedoch um einen sehr begrenzten Lehrzeitraum, an dessen Ende die Erfolgskontrolle auch mittels einer praktischen Prüfung erfolgen soll. Deshalb muss eine praktische Prüfung auch mit den gegebenen Ressourcen durchführbar sein.

Im Allgemeinen werden in einer OSCE Ärzte der jeweiligen Abteilung als Prüfer eingesetzt. Nach Newble et al. besteht dabei kein Unterschied zwischen Prüfern, die ein Training vor der OSCE erhielten, gegenüber denen, die ohne Training bewerteten. Allerdings zeigt er eine deutliche Verbesserung der Reliabilität, wenn inkonsistente Prüfer von der Prüfung ausgeschlossen werden (Newble et al. 1980). Zur Vorbereitung der Prüfer wird eine Zeit von 30 Minuten direkt vor der Prüfung empfohlen, um sich mit der Prüfungsumgebung und der Checkliste vertraut zu machen (O'Connor und McGraw 1997). Einige Autoren setzen auch die Simulationspatienten für die Bewertung der eigenen Station ein. Dabei zeigen sich niedrige bis gute Werte im Bereich Reliabilität und Validität (McLaughlin et al. 2006; Wilkinson und Fontaine 2002). Chenot et al. zeigt, dass gut geschulte Studenten höherer Semester als Prüfer in OSCE-Prüfungen eingesetzt werden können (Chenot et al. 2007).

1.4 Qualitätskriterien einer klinisch- praktischen Prüfung

Prüfungen in der medizinischen Ausbildung haben neben ihrer Relevanz für die Studierenden und die Universitäten auch direkte gesellschaftliche Bedeutung, deshalb werden an Prüfungen bestimmte Anforderungen hinsichtlich ihrer Qualität als Messinstrument gestellt. Die Qualität eines Tests bzw. eines Fragebogens lässt sich an drei zentralen Kriterien der Testgüte festmachen: Objektivität, Reliabilität und Validität (Bloch et al. 1999; Bortz und Döring 2002; Schumacher und Brähler 2006).

Die Objektivität eines Tests gibt an, in welchem Ausmaß die Testergebnisse vom Testanwender unabhängig sind. Vor allem durch eine klare Strukturierung

und Standardisierung eines Tests, z.B. durch exakten Wortlaut von Instruktionen und das Verwenden von Auswerteschablonen, wird eine hohe Objektivität erreicht. Die numerische Bestimmung der Objektivität eines Tests erfolgt über die durchschnittliche Korrelation der Ergebnisse verschiedener Testanwender. Wenn diese Korrelation nahe 1 liegt, kann Objektivität vorausgesetzt werden.

Die Reliabilität eines Tests kennzeichnet den Grad der Genauigkeit, mit dem das geprüfte Merkmal gemessen wird. Bei einer hohen Reliabilität wird eine Wiederholung der Prüfung weitgehend zu den gleichen Resultaten führen. Die Korrelationskoeffizienten können dabei zwischen 0 und 1 liegen. Als Richtgröße für eine sehr gute Reliabilität kann ein Wert über 0,80 gelten. Die Messzuverlässigkeit wird durch Einflüsse beeinträchtigt, die nichts mit dem zu tun haben, was die Prüfung messen soll. Dazu gehören neben einer mangelhaften Objektivität etwa auch Rateeinflüsse oder sprachliche Missverständnisse. Als wichtigster Störfaktor der Reliabilität wurde in neuerer Zeit die Problem- oder Fallspezifität erkannt (van der Vleuten 1996). Die Kandidatenleistung variiert über verschiedene Probleme und Patientenfälle stark. Erst anhand der Leistungen in einer genügend großen Stichprobe von Fällen und Problemen lässt sich zuverlässig aussagen, wie ausgeprägt die Kompetenz eines Kandidaten ist.

Die Validität eines Tests gibt an, wie gut der Test in der Lage ist, genau das zu messen, was er zu messen vorgibt. Es wird in drei Hauptarten von Validität unterschieden: Inhaltsvalidität, Kriteriumsvalidität und Konstruktvalidität.

Inhaltsvalidität: Inhaltsvalidität (Face Validity, Augenscheinvalidität, Logische Validität) ist gegeben, wenn der Inhalt der Test- Items das zu messende Konstrukt in seinen wichtigsten Aspekten erschöpfend erfasst. Die Höhe der Inhaltsvalidität eines Tests kann nicht numerisch bestimmt werden, sondern beruht allein auf subjektiven Einschätzungen. Streng genommen handelt es sich bei der Inhaltsvalidität deswegen auch nicht um ein Testgütekriterium, sondern nur um eine Zielvorgabe, die bei der Testkonstruktion bedacht werden sollte.

Kriteriumsvalidität: Kriteriumsvalidität (kriterienbezogene Validität) liegt vor, wenn das Ergebnis eines Tests zur Messung eines latenten Merkmals bzw. Konstrukts (z.B. Berufseignung) mit Messungen eines korrespondierenden manifesten Merkmals bzw. Kriteriums übereinstimmt. Die Kriteriumsvalidität ist

definiert als Korrelation zwischen den Testwerten und den Kriteriumswerten einer Stichprobe. Leider ist die Kriteriumsvalidierung in ihrem Anwendungsbereich dadurch stark eingeschränkt, dass vielfach kein adäquates Außenkriterium benannt werden kann. Neben der Schwierigkeit, überhaupt ein angemessenes Außenkriterium zu finden, stellt sich auch die Frage nach der Operationalisierung des Kriteriums. Sind Kriteriumswerte invalide oder unreliabel erfasst, so ist natürlich jede Validierung mit diesem Kriterium unbrauchbar.

Konstruktvalidität*:* Der Konstruktvalidität kommt besondere Bedeutung zu, da Inhaltsvalidität kein objektivierbarer Kennwert ist und Kriteriumsvalidierung nur bei geeigneten Außenkriterien sinnvoll ist.

1.5 Zielsetzung

Mit der Einführung der neuen ÄAppO wird mehr praktisch orientierte Lehre gefordert. Die OSCE eignet sich, diese erworbenen praktischen Fertigkeiten zu beurteilen. Bislang liegen in Deutschland, wie auch international, nur wenige Erfahrungen mit einer klinisch-praktischen (OSCE-) Prüfung im Bereich der Notfallmedizin für Studierende vor.

In Anbetracht der erheblichen Anforderungen an eine klinisch-praktische Prüfung einerseits und der limitierten Ressourcen anderseits, ist es Ziel dieser Arbeit:

- die Planung, Entwicklung und Durchführung einer validen praktisch-klinischen Prüfung im Querschnittsfach Notfallmedizin zu beschreiben;
- die Reliabilität mittels der Videoauswertung zu bestimmen;
- studentische mit ärztlichen Ratern im Videorating zu vergleichen;
- die Itemschwierigkeit und Trennschärfe zu errechnen;
- die Objektivität mittels zuvor trainierter Teststudenten zu beurteilen.

2. MATERIAL und METHODEN

2.1 Entwicklung einer klinisch- praktischen Prüfung

2.1.1 Beteiligte Personen und Aufgaben

Die wichtigste Personengruppe, die maßgeblich an der Entwicklung der Prüfung beteiligt ist, ist die Prüfungskommission (Krebs 1999). Sie ist verantwortlich für die

- Festlegung des Blueprints und der Prüfungsziele.
- Planung der Prüfungsentwicklung.
- Rekrutierung und Instruktion der Autoren und Examinatoren.
- Benennung des Durchführungsverantwortlichen.
- Revision der erstellten Fälle.
- Auswahl der Stationen für eine Prüfung und Standardsetzung.
- Erstellung der schriftlichen Prüfungsunterlagen und Anmeldungsadministration.
- Prüfungsauswertung und Prüfungsevaluation.

2.1.2 Festlegung der Prüfungsziele und –stationen

Als Prüfungsziele, die sinnvollerweise einerseits im Modul praktisch unterrichtet und andererseits in einer praktischen Prüfung evaluiert werden können, wurden hinsichtlich der Diagnostik und Behandlung eines notfallmedizinischen Patienten definiert:

1. Erkennen und Erfassen der Vitalparameter;
2. Einschätzung des Schweregrades der Erkrankung;
3. Sicherung der Atemwege mit Basis-Hilfsmitteln;
4. Diagnostik und Therapie der kardialen und Kreislaufsituation mit Hilfsmitteln;
5. Begleitende Maßnahmen bei der Versorgung kritisch kranker Patienten;

6. Diagnostik und Behandlung unter Berücksichtigung der sicherheitsrelevanten Aspekte für den Patienten, den Arzt und beteiligte Personen.

Unter Berücksichtigung der limitierten personellen, materiellen und zeitlichen Ressourcen des notfallmedizinischen Moduls, hat sich die Prüfungskommission auf zunächst fünf praktisch-klinische Prüfungsstationen mit je neun Messitems geeinigt. Die OSCE wurde als kompensatorische Prüfungsform eingesetzt, so dass der Prüfling lediglich eine bestimmte Gesamtpunktzahl erreichen muss und nicht zufriedenstellende Leistungen an einer Station mit sehr guten Leistungen an einer anderen Station ausgleichen kann (Newble 2004). Weiterhin wurden die neun Items inhaltlich und zeitlich in drei Abschnitte mit je zwei bis vier Items weiterhin unterteilt, so dass für die Studierenden die Möglichkeit bestand, mit der Prüfung fortzufahren, auch wenn nicht alle Items erfolgreich bestanden wurden. Die Prüfungskommission war sich dabei den Anforderungen an die klassischen OSCE- Prüfungen bewusst, dass eine repräsentative und damit hinreichend zuverlässige Prüfung in der Regel eine genügend große Zahl von Stationen erfordert, die sowohl inhaltlich verschieden sind als auch unterschiedliche Kompetenzen evaluieren (Bloch et al. 1999; Krebs 1999; Newble und Swanson 1988). Die hier evaluierten praktisch-klinischen Kompetenzen sind aber einerseits begrenzt und andererseits wurde zugleich eine schriftliche Prüfung in Form einer Multiple-Choice-Question(MCQ)-Klausur eingefügt, um die geforderte Heterogenität der Prüfung zu gewährleisten. Die inhaltliche Grundlage bildeten, die zum Zeitpunkt der Untersuchung geltenden Empfehlungen des European Resuscitation Councils (European Resuscitation Council 2000).

Um die Validität der Prüfung sicherzustellen, wurde die Abbildung der Prüfungsziele auf die OSCE-Stationen mittels eines Blueprintes entwickelt und dargestellt. Ein Blueprint ist ein gewichtetes Verzeichnis der Prüfungsinhalte. Die Autoren wurden bestimmt und damit beauftragt, unter Berücksichtigung der Lehrinhalte und des Blueprints, eine Prüfungsstation zu erstellen. Alle Autoren der Prüfungsbögen waren auch als Prüfer in der OSCE beteiligt. Die Bögen wurden in der Revision von der Prüfungskommission bearbeitet und verabschiedet.

Tab. 2.1 Blueprint

Prüfungsstation / Prüfungsziel	BLS	Defi	Rhythmus	Airway	Trauma	**Gewichtung**
1 Vitalparameter einschätzen	2	2	2	2	4	**12**
2 Schweregrad bestimmen	2	0	4	2	4	**12**
3 Sicherung von O_2 und Ventilation	6	0	0	12	2	**20**
4 Kardiozirkulatorische Therapie	6	14	10	0	2	**32**
5 Begleitende Maßnahmen	2	0	2	2	6	**12**
6 Sicherheit	2	4	2	2	2	**12**
Anteil an der Prüfung	**20**	**20**	**20**	**20**	**20**	**100**

Anteile der Prüfungsziele an der jeweiligen Prüfungsstation in [%]

2.2 Planung der OSCE-Prüfung

Die Aufgaben für Planung und Durchführung beinhalteten das Finden und Reservieren von geeigneten Prüfungslokalitäten, die Planung, Beschilderung und den Aufbau des Prüfungscircuits, Beschaffung der erforderlichen Einrichtungs- und Prüfungsmaterialien, der Erstellung von Zeitplänen und Checklisten für die Materialkontrolle, der Organisation der Prüfer-Raumpläne, der Pausenregelung, die Verpflegung und die Bereitstellung eines „Troubleshouters“ für logistische und inhaltliche Fragen. Insgesamt wurden zwei identische Prüfungsparcours angelegt, um die Studierenden eines gesamten Semesters an einem Tag prüfen zu können.

Zur Erleichterung der Prüfungsbewertung, der Auswertung und auch als Beitrag zur Objektivierung der Prüfung wurden in Zusammenarbeit mit dem Bereich Informationstechnologie maschinenlesbare Bögen entwickelt. Aus statistischen und organisatorischen Gründen wurde die tatsächlich benötigte Prüfungsdauer zusätzlich dokumentiert. Ebenso bestand für die Prüfer die Möglichkeit, den Studierenden unabhängig von der Checkliste mit einer globalen Gesamtbeurteilung zu bewerten (Schulnotenskala sehr gut (1) bis ungenügend (6)). Hierbei sollte der Gesamteindruck des Studierenden im Hinblick auf die

gestellten Prüfungsfragen beurteilt werden. Die Globalbeurteilung ging nicht in die Modulnote der Studenten ein.

Neben der Instruktion der Unterrichtsdozenten des Kleingruppenunterrichts unter Berücksichtigung der Prüfungsziele wurden nochmals alle Prüfungsdozenten, die sich größtenteils aus den Unterrichtsdozenten rekrutierten, in die Inhalte, Organisation und Ablauf der Prüfung eingewiesen. Ein exemplarischer Durchlauf wurde geprobt, um Unsicherheiten im Umgang mit den Prüfungsbögen zu beseitigen und Fragen zu klären.

2.3 Training der Teststudierenden

Schon im Wintersemester 2004/2005 wurden acht Teststudierende aus den Absolventen des Notfallmedizinischen Moduls des vorherigen Semesters rekrutiert, die bereits einen Pilot-OSCE ohne Bewertung durchlaufen hatten. Diese wurden in je zwei Gruppen zu vier Studierenden aufgeteilt:

Gruppe A wurde auf Bestehen jeder Prüfungsstation mit vollen neun Punkten trainiert, Gruppe B auf sechs Punkte pro Prüfung. Bei den Teststudenten der Gruppe B wurden somit drei definierte, gleich bleibende Fehler eingebaut. Die Teststudierenden hatten an je zwei Abenden unter Ausschluss der Unterrichts- und Prüfungsdozenten die Möglichkeit, die erwartete Prüfungsnote zu trainieren. Das Training und die Einschleusung in den „echten“ Prüfungsparcours waren nur dem Durchführungsverantwortlichen, dem Abteilungsleiter und den wissenschaftlichen Hilfskräften bekannt, die nicht an der Bewertung der OSCE-Prüfung teilnahmen. Die Gruppen der Teststudierenden durchliefen jeweils beide Prüfungsparcours.

2.4 Durchführung der OSCE-Prüfung

Die Prüfung fand am Ende des notfallmedizinischen Moduls im Sommersemester 2005 ganztätig statt. Die Themen der Stationen lauteten: „Versorgung eines verunfallten Patienten“ (Trauma), „Basic-Life-Support (BLS)“, „Rhythmusdiagnostik“ (Rhythmus), „Sicherung der Atemwege“ (Airway) und

„Defibrillation“ (Defi). Im Lehr- und Simulationszentrum für Anästhesiologie, Rettungs- und Intensivmedizin wurden zwei identische Prüfungsparcours aufgebaut (Anlage 6.1). Jeder Prüfungsraum war mit einem zuvor geschulten Prüfer besetzt, der sowohl für die Kommunikation mit den Studierenden (Begrüßung, Aufgabenstellung, ggf. Interventionen) zuständig war, wie auch für die Bewertung der Leistung mithilfe der Checkliste und anhand einer Globalbeurteilung auf den Prüfungsbögen. Die Prüfer evaluierten jeweils 90 Minuten, gefolgt von einer 20minütigen Pause. Danach wechselten Sie die Prüfungsstation. Die kalkulierte Prüfungsdauer mit je zwei gleichartig aufgebauten Prüfungsstationen bei 176 Studierenden betrug acht Stunden. Insgesamt 12 Prüfer (zwei Prüfer zur Pausenauslösung), zwei wissenschaftliche Hilfskräfte und der Durchführungsverantwortliche wurden für die Prüfung eingeplant. Am Prüfungstag wurden die Studierenden nach einem festen Zeitplan in Gruppen einbestellt (Anlage 6.2) und nach der Registrierung mit einer Identifikationskarte (Matrikelnummer und Prüfungsnummer) einzeln, in Abständen von fünf Minuten, an die erste Station geschickt. Pro Station hatten die Studenten 4,5 Minuten zur Lösung der Aufgabe Zeit. 0,5 Minuten wurden für die Studierenden als Übergangszeit zur nächsten Station und für Prüfer zum Eintragen der Daten in den Kopfteil des Prüfungsbogens veranschlagt. Mit Hilfe einer Stoppuhr war jeder Prüfer selbst für die Kontrolle der Wechselzeit zuständig.

2.5 Die Videoauswertung

2.5.1 Gewinnen des Videomaterials

Für die im Sommersemester 2005 durchgeführte Videountersuchung wurde nur der Parcours A gewählt. In den Prüfungsräumen wurde je eine Videokamera aufgestellt, die den Prüfungsraum während der gesamten OSCE auf Mini-DV bzw. VHS Kassette aufnahm. Die Prüflinge wurden über die Videoaufzeichnung und die spätere wissenschaftliche Verwendung des Materials aufgeklärt und gaben ihr Einverständnis mittels Unterschrift. Kandidaten, die einer Aufnahme ihrer Prüfung nicht zustimmten, konnten, ohne dass Ihnen ein Nachteil entstand, in den Parcours B ohne Videoaufzeichnung wechseln.

Zu Beginn jeder einzelnen Prüfungsstation wurde vom Prüfer die Identifikationskarte des Kandidaten vor die Kamera gehalten, um die Wiedererkennung der Studenten bei der späteren Videoauswertung zu erleichtern.

2.5.2 Das Videorating

Für das Videorating wurden vier Rating-Gruppen gebildet. Zwei Studentengruppen (Stud 1, Stud 2) , die aus Medizinstudenten im klinischen Studienabschnitt bestanden, und zwei Gruppen mit ärztlichen Ratern, wobei eine Gruppe aus Fachärzten der Anästhesie (Fachärzte) und die andere Gruppe aus Assistenzärzten in der Weiterbildung der Anästhesie (Assistenten) bestand. Personell konnten diese Gruppen variieren. Alle Rater bekamen eine Einführung, bei der alle Stationen, der Prüfungsablauf und die Bewertungsbögen mit dem Identifikationsteil und den Checklisten vorgestellt wurden. Dann hatten die Rater vier Wochen Zeit, um zu Hause das Videorating durchzuführen.

2.6 Statistische Methoden

2.6.1 Daten- Auswertung

Die Prüfungsbögen wurden mithilfe der Abteilung für Informationstechnologie des Bereichs Humanmedizin der Universität Göttingen maschinell eingelesen. Die statistischen Berechnungen erfolgten mithilfe von SPSS 11.5 für Windows und Microsoft Office Excel für Windows. Die Berechnung der Kappa-Koeffizienten wurde durch die Abteilung Medizinische Statistik des Bereiches Humanmedizin der Universität Göttingen mit SAS durchgeführt.

2.6.2 Reliabilität

Für das Maß der Reliabilität wurde mit den Ergebnissen aus OSCE-Prüfung und Videorating Cohens κ (kappa) errechnet. Cohens κ eignet sich zur Berechnung eines zufallskorrigierten Übereinstimmungsmaßes (Wirtz und Caspar 2002).

Für die einzelnen Prüfungstationen wurden der gewichtete Kappa-Koeffizient berechnet. Während bei der Berechnung des ungewichteten Kappa-Koeffizienten nur vollständig übereinstimmende Werte berücksichtigt werden, so wird beim gewichteten Kappa-Koeffizienten die Abweichung der einzelnen Antworten mit berücksichtigt; eine teilweise Übereinstimmung kann somit besser erfasst werden als mit dem ungewichteten Koeffizienten (Cohen J. 1960; Cohen J. 1968). Ein Kappa-Koeffizient von 0-0,20 gilt als „schlechte“, von 0,21-0,40 als „ausreichende“, von 0,41-0,60 als „moderate“, von 0,61-0,80 als „gute“ und von über 0,81 als „hervorragende“ Übereinstimmung (Brennan und Silman 1992).

Da am Videorating mehrere Ratergruppen teilgenommen haben, wurden Raterpaare gebildet, wobei jeweils der Originalprüfungsrater einer Videoratergruppe gegenübergestellt wurde. Der Kappa-Koeffizient wurde für alle Raterpaare ermittelt und der Median dieser Werte als Schätzung der durchschnittlichen Übereinstimmung zwischen allen Ratern betrachtet (Roth 1984).

2.6.3. Vergleich der Ratergruppen

Zum Vergleich der Videoratergruppen mit dem Originalrater wurde der Spearman Rangkorrelationskoeffizient r_s berechnet. Der Spearman Rangkorrelationskoeffizient r_s eignet sich zur Messung eines Zusammenhangs zweier ordinalskalierter Merkmale, wenn keine Normalverteilung gegeben ist. Der Koeffizient kann Werte zwischen -1 bis 1 annehmen. Dabei deutet -1 auf einen maximal gegensinnigen, monotonen Zusammenhang der Merkmale hin und 1 auf einen maximal gleichsinnigen, monotonen Zusammenhang. Sind die Merkmale unabhängig, erhält man einen Korrelationskoeffizienten von 0. Ein deutlicher Zusammenhang zweier Merkmale wird ab einem Koeffizienten von >0,5 angenommen (Bortz und Lienert 1998).

2.6.4 Itemschwierigkeit und Trennschärfe

Die Itemschwierigkeit gibt die Anzahl richtiger Lösungen durch die Gesamtzahl der Antworten auf ein Item an. Die Itemschwierigkeit ist 0, wenn niemand die Aufgabe lösen kann und 1 wenn alle Probanden richtig antworten. Die Itemschwierigkeit ist keine Eigenschaft des Items ansich, sondern ist immer im Bezug auf die Prüfunspopulation zu sehen. Die Schwierigkeit eines Items hat Auswirkungen auf die potentielle Trennschärfe und die Reliabilität einer Prüfung. Die Itemschwierigkeit sollte über 20 liegen, da bei niedrigerer Itemschwierigkeit der Einfluss des Ratens zu groß wird. Items im Schwierigkeitsbereich von 41-95 weisen mit großer Wahrscheinlichkeit eine gute Trennschärfe auf. Die einzelnen Unterpunkte pro Prüfungsstation wurden als dichotome Items betrachtet und die Itemschwierigkeit als arithmetisches Mittel errechnet.

Die Trennschärfe zeigt an, wie gut die einzelne Aufgabe die Probanden mit einem „gutem“ Testergebnis von denen mit einem „schlechten“ Testergebnis trennt. Die Trennschärfe kann Werte zwischen -1 und 1 annehmen. Eine hohe Trennschärfe eines Items gibt an, dass Kandidaten, die diese Aufgabe lösen, auch im Gesamttest gut abschneiden. Eine Trennschärfe um null zeigt, dass das Item von „guten“ wie auch von „schlechten“ Probanden gleichermaßen beantwortet wird. Eine negative Trennschärfe bringt zum Ausdruck, dass „schlechte“ Probanden das Item richtig und „gute“ Probanden es falsch beantworten. Items mit negativer Trennschärfe sollten vermieden werden. Eine adäquate Trennschärfe liegt über 0,20. Eine niedrige Trennschärfe liegt zwischen 0,1-0,19, Items zwischen 0,09- -0,09 weisen keine Trennschärfe auf. Unter 0,09 liegt eine negative Trennschärfe vor. Bei dichotomen Items wird die Trennschärfe als punktbiserale Korrelation berechnet (Krebs 1999; Lienert et al. 1998).

3. Ergebnisse

3.1 Prüfungsergebnisse

An der OSCE Prüfung im Sommersemester 2005 nahmen 176 Studierende aus dem 5. klinischen Semester teil. In die Videoanalyse gingen 91 Studenten ein. Bei fünf Prüfungstationen mit jeweils neun Items gab es eine Gesamtpunktzahl von 45 zu erreichen. Die absolute und prozentuale Notenverteilung ist Tabelle 3.1 zu entnehmen. In Tabelle 3.2 werden die Prüfungsergebnisse als absolute Punktzahl und Notenmittelwert der Studierenden für die jeweilige Prüfungsstation und insgesamt dargestellt. Zusätzlich werden die Ergebnisse zeitlich getrennt dargestellt, nach Absolvierung der ersten und zweiten Hälfte der Studierenden, die absoluten Differenzen errechnet.

Tab. 3.1 Notenverteilung der OSCE – Prüfung

Note	Gesamtpunkte	[n]	%
1	41- 45	29	16,48
2	36- 40	97	55,11
3	32- 35	37	21,02
4	27- 31	12	6,81
5	23- 26	1	0,57
6	0-22	0	0

[n]: Anzahl der Studenten; %: prozentuale Häufigkeit

Tab. 3.2. Bewertung der Prüfungsstationen; 1. und 2. Hälfte

Prüfung		Gesamt (Mittel ± SD) n = 176	1.Hälfte (Mittel ± SD) n = 88	2.Hälfte (Mittel ± SD) n = 88	Differenz
BLS	Pkt.	7,75±1,19	7,90±1,09	7,60±1,28	-0,3
	Note	1,89±0,85	1,71±0,78	2,07±0,88	0,36
Defi	Pkt.	7,82±0,99	7,83±1,06	7,82±0,92	-0,01
	Note	2,28±0,91	2,15±0,89	2,41±0,92	0,26
Rhythmus	Pkt.	7,66±1,30	7,77±1,21	7,55±1,38	-0,22
	Note	2,52±0,98	2,43±0,92	2,61±1,03	0,11
Airway	Pkt.	7,28±1,23	7,30±1,17	7,26±1,30	-0,04
	Note	2,37±0,85	2,45±0,76	2,30±0,94	-0,14
Trauma	Pkt.	6,88±1,44	6,94±1,45	6,82±1,44	-0,12
	Note	2,58±0,98	2,40±0,91	2,76±1,02	0,36
Gesamt	Pkt.	37,18±3,51	37,74±3,42	37,05±3,60	-0,69
	Note	2,32±0,70	2,25±0,68	2,39±0,72	0,14

Mittel: arithmetisches Mittel ; SD: Standardabweichung; n: Anzahl der Studenten; Pkt: Punkte

Die Tabellen 3.3-3.7 zeigen die einzelnen Prüfungsstationen mit der prozentualen Häufigkeit der jeweiligen korrekt erfüllten Checklistenitems. Verglichen werden die beiden Prüfungsparcours, von denen der Parcours A anschließend von Videoratern bewertet wurde, während der Parcours B nicht in die Videobewertung einging. 15 der 45 Checklistenitems weisen einen signifikanten Unterschied zwischen den beiden Parcours auf.

Tab. 3.3 Bewertung BLS: Vergleich Parcours A und B

BLS-Prüfung Prüfungsitem	Parcours A Korrekt %	Parcours B Korrekt %	p
1.1 Eigensicherung	74,7	88,2	ns
1.2 Bewusstsein prüfen	100	100	ns
1.3 Hilferuf	67	83,5	<0,05
1.4 "Atemwege frei machen"	81,3	95,2	<0,05
1.5 Notruf	69,2	89,4	< 0,001
1.6 2 effektive Atemzüge	93,4	91,8	ns
1.7 Puls prüfen	93,4	87,7	ns
1.8 Druckpunkt aufsuchen	59,3	89,5	< 0,001
1.9 HDM 15:2	95,6	96,5	ns

Korrekte Checklistenitems der Parcours A (Video) und B (nicht Video); ns: nicht signifikant

Tab. 3.4 Bewertung Defi: Vergleich Parcours A und B

Defi-Prüfung Prüfungsitem	Parcours A Korrekt %	Parcours B Korrekt %	p
2.1 Einschalten, Gel auf Thorax	80	86,8	ns
2.2 Positionierung	98,9	99	ns
2.3 Rhythmusanalyse	85,9	95,6	<0,05
2.4 Ansage „Achtung, ich lade“	95,3	87,9	ns
2.5 Ladung 200J, Paddels auf Thorax	100	95,6	ns
2.6 „Alles weg vom Patienten“ + Kontrolle	91,7	79,1	<0,05
2.7 „Achtung Schock“	97,6	99	ns
2.8 weitere Schocks mit 200J und 360J	95,3	91,2	ns
2.9 Paddels zurück, Puls und RR	78,8	9,9	<0,001

Korrekte Checklistenitems der Parcours A (Video) und B (nicht Video); ns: nicht signifikant

Tab. 3.5 Bewertung Rhythmus: Vergleich Parcours A und B

Rhythmus-Prüfung Prüfungsitem	Parcours A Korrekt %	Parcours B Korrekt %	p
3.1 Kabel anbringen	90,1	93	ns
3.2 Optimierung Amplitude	80,2	81,2	ns
3.3 Kammerfrequenz: tachykard/bradykard	94,5	100	<0,05
3.4 Kammerfrequenz: regelmäßig/unregelmäßig	89	97,6	<0,05
3.5 QRS-Komplex: schmal/breit	92,3	96,5	ns
3.6 korrekte Bezeichnung	84,6	85,9	ns
3.7 Vitalzeichenkontrolle	66	62,4	ns
3.8 Therapie BLS	75,8	65,9	ns
3.9 Therapie Adrenalin, Atropin o. trankutaner SM	92,3	74,1	<0,001

Korrekte Checklistenitems der Parcours A (Video) und B (nicht Video); ns: nicht signifikant

Tab. 3.6 Bewertung Airway: Vergleich Parcours A und B

Airway-Prüfung Prüfungsitem	Parcours A Korrekt %	Parcours B Korrekt %	p
4.1. O2 Gabe	87,9	98,8	<0,05
4.2 "Freimachen der Atemwege"	84,6	90,6	ns
4.3 Hilfe holen	90,1	80	ns
4.4. Maskenbeatmung	90,1	97,6	<0,05
4.5 Mit Reservoir oder Demandventil	64,8	64,7	ns
4.6 Laryngoskop mit Funktionkontrolle	34,1	96,5	<0,001
4.7. Tubus, Führungsstab und Blockerspritze	98,9	98,8	ns
4.8. Absaugung und Absaugkatheter	55	61,2	ns
4.9. Stethoskop, CO2 Kontrolle	89	75,3	<0,05

Korrekte Checklistenitems der Parcours A (Video) und B (nicht Video); ns: nicht signifikant

Tab. 3.7 Bewertung Trauma: Vergleich Parcours A und B

Trauma- Prüfung Prüfungsitem	Parcours A Korrekt %	Parcours B Korrekt %	p
5.1 Ansprechen, Vorstellen, Vigilanzkontrolle	82,4	72,9	ns
5.2 Atmungskontrolle	49,5	62,4	ns
5.3 Kreislaufkontrolle	62,6	51,8	ns
5.4 Bodycheck: Kopf/Hals	94,5	85,9	ns
5.5 Bodycheck: Stamm	90,1	91,8	ns
5.6 Bodycheck: Extremitäten	50,6	64,7	ns
5.7 Maßnahmen 1+2	98,9	100	ns
5.8 Maßnahmen 3+4	91,2	97,6	ns
5.9 Maßnahmen 5+6	53,9	76,5	<0,05

Korrekte Checklistenitems der Parcours A (Video) und B (nicht Video); ns: nicht signifikant

3.2 Reliabilität

3.2.1 Kappa-Koeffizient - Checklistenbewertung

Tabelle 3.8 zeigt die Kappa-Koeffizienten κ nach Cohen für die Gesamtprüfung sowie für die Einzelstationen mittels Checkliste. Ein κ-Wert entspricht einem Raterpaar bestehend aus dem Originalrater gegenüber einem Videorater (Stud 1, Stud 2, Fachärzte, Assistenten). Cohens Kappa für die Gesamtprüfung wird auf 0,64-0,74 berechnet. Der Median liegt bei 0,73. Der Kappa-Koeffizient für die Stationen liegt zwischen 0,37-0,74. Als Median vom κ der einzelnen Stationen ergeben sich Werte zwischen 0,44-0,69 ($p<0,001$).

Tab 3.8. Kappa- Koeffizienten der Checklistenbewertung;

	Checklistenbewertung				
Original	**Stud 1**	**Stud 2**	**Fachärzte**	**Assistenten**	**Median**
Gesamt	0,64	0,72	0,74	0,73	0,73
BLS	0,74	0,64	0,73	0,62	0,69
Defi	0,6	0,55	0,45	0,5	0,53
Rhythmus	0,69	0,68	0,55	0,66	0,67
Airway	0,66	0,63	0,37	0,68	0,65
Trauma	0,37	0,57	0,4	0,48	0,44

3.2.2 Kappa-Koeffizient - Globalbewertung

Die Tabelle 3.9 zeigt den Kappa- Koeffizienten nach Cohen für die einzelnen Prüfungsstationen beim Betrachten der Note aus der Globalbewertung. Kappa liegt zwischen 0,23-0,57. Der Median ist im Bereich 0,33-0,46 zu finden ($p<0,001$). Da keine Gesamtnote in der OSCE vergeben wurde, konnte auch kein Kappa-Koeffizient für eine Gesamtnote berechnet werden.

Tab. 3.9 Kappa- Koeffizient der Globalbewertung

	Globalbewertung				
Original	**Stud 1**	**Stud 2**	**Fachärzte**	**Assistenten**	**Median**
BLS	0,55	0,41	0,23	0,38	0,4
Defi	0,39	0,45	0,23	0,26	0,33
Rhythmus	0,56	0,44	0,23	0,42	0,43
Airway	0,45	0,46	0,3	0,48	0,46
Trauma	0,34	0,49	0,26	0,38	0,36

3.3 Vergleich studentische und ärztliche Prüfer

3.3.1 Vergleich bei Checklistenbewertung

In Tabelle 3.10 sind die arithmetischen Mittel der Punkte der einzelnen Stationen in der Originalprüfung sowie der studentischen und ärztlichen Videorater dargestellt. Für den Vergleich wurden die zwei Gruppen von studentischen bzw. ärztlichen Prüfern zu jeweils einer Gruppe zusammengefasst. An den Stationen BLS und Defi bewerteten studentische wie auch ärztliche Prüfer besser als in der Originalprüfung. An der Station-Trauma wurden von den Videoratern weniger Punkte vergeben als von den Originalprüfern. An der Rhythmusstation bewerteten die Studenten besser und die Ärzte schlechter als in der Originalprüfung. Dagegen bewerteten die Ärzte in der Airway-Station besser und die Studenten vergaben weniger Punkte als die Originalprüfer. Die maximale Differenz liegt bei den Studenten zwischen -0,51-0,11, bei den Ärzten zwischen -0,48-0,25. Bei den Gesamtpunkten zeigt sich, dass die Studenten im Durchschnitt schlechter bewertet haben als die Originalprüfer und die ärztliche Videorater insgesamt etwas mehr Punkte vergeben haben.

Tab. 3.10 Vergleich Studenten – Ärzte bei Checklistenbeurteilung

	Checklistenbewertung		
	Original	**Studenten**	**Ärzte**
BLS	7,34	7,45	7,57
Defi	7,44	7,46	7,69
Rhythmus	7,65	7,74	7,32
Airway	6,95	6,64	7,4
Trauma	6,74	6,23	6,26
Gesamt	36,11	35,52	36,24

arithmetisches Mittel der Punkte bei Checklistenbeurteilung

Die Punktedifferenz veranschaulicht Abbildung 3.1. Hier ist die Punktedifferenz der Gesamtpunktzahl im Videorating gegen die Anzahl der Prüflinge aufgetragen. Deutlich ist zu erkennen, dass die Studenten eher weniger Punkte vergeben haben als die Ärzte. So sind von den Studenten 53 Prüflinge schlechter und 29 Prüflinge besser bewertet worden. Bei den Ärzten sind 37

Kandidaten mit weniger Punkten und 48 Kandidaten mit mehr Punkten bewertet worden. Die gleiche Punktzahl sowohl in der Originalprüfung wie auch im Videorating erhielten von den Studenten neun und von den Ärzten sechs Prüflinge.

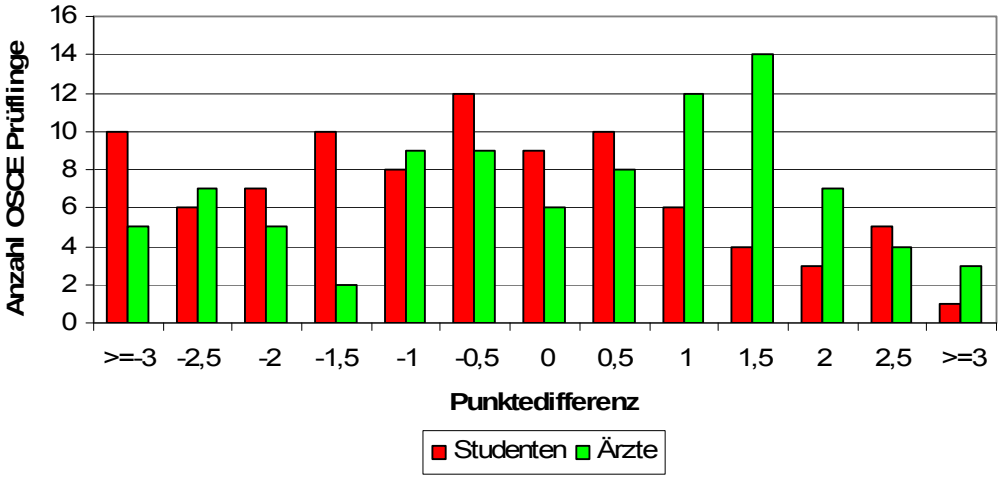

Abb. 3.1 Punktedifferenz: Studenten – Ärzte bei Checklistenbewertung

In der Abbildung 3.2 sind die Spearman-Rangkorrelationskoeffizienten r_s für die Checklistenbewertung der Videorater aufgetragen. Der Korrelationskoeffizient ist bis auf die Airway-Station bei den studentischen Prüfern höher als bei den ärztlichen Prüfern.

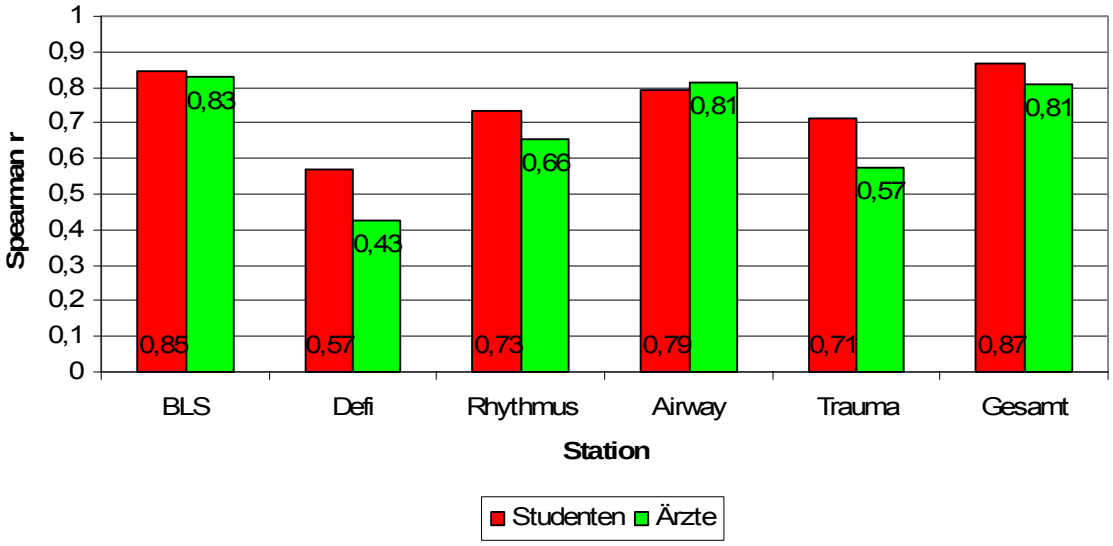

Abb. 3.2 Spearman r_s: Studenten – Ärzte bei Checklistenbeurteilung

3.3.2 Vergleich bei Globalbeurteilung

In Tabelle 3.11 sind die Notenmittelwerte der einzelnen Stationen von Original- und Videoratern an allen Stationen bei Globalbeurteilung abgebildet. Bessere Noten wurden von beiden Videoratergruppen an der Airway- wie auch an der Defi-Station vergeben. Die Studenten bewerteten außerdem an der Rhythmus-Station besser als die Originalprüfer. Die Ärzte vergaben an dieser Station schlechtere Noten. Ärzte wie auch Studenten bewerteten die Kandidaten an den Stationen BLS und Trauma mit schlechteren Noten als in der Originalprüfung. Die maximale Differenz liegt bei Studenten zwischen -0,14-0,21, bei den Ärzten zwischen -0,64-0,66. Insgesamt haben die Ärzte schlechtere Noten vergeben, wohingegen die Studenten die Prüflinge ein wenig besser benotet haben.

Tab. 3.11 Vergleich Studenten – Ärzte bei Globalbeurteilung

	Globalbeurteilung		
	Original	Studenten	Ärzte
BLS	1,95	2,03	2,61
Defi	2,48	2,34	1,84
Rhythmus	2,44	2,37	2,76
Airway	2,38	2,27	2,03
Trauma	2,47	2,68	2,69

arithmetisches Mittel der Noten bei Globalbeurteilung

In Abbildung 3.3 sind die Spearman-Rangkorrelationskoeffizienten r_s jeder Station von studentischen und ärztlichen Ratern aufgetragen. Mit Werten zwischen 0,57-0,75 liegen die studentischen Rater an vier der fünf Stationen über den Koeffizienten der ärztlichen Rater, die zwischen 0,48-0,67 liegen. An der Airwaystation ist r_s bei den Ärzten höher als bei den Studenten.

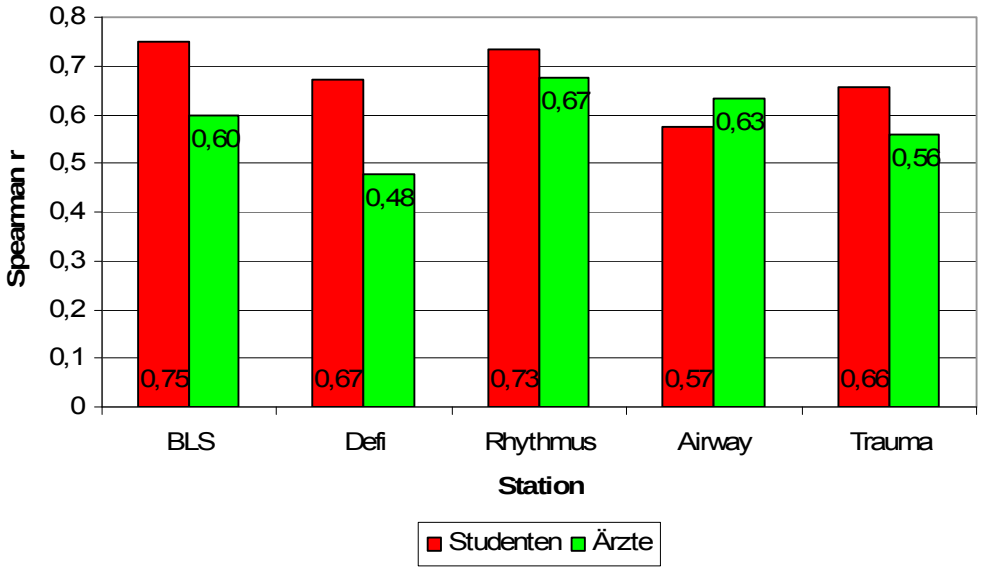

Abb. 3.3 Spearmans r_s: Studenten – Ärzte bei Globalbeurteilung

3.4 Itemschwierigkeit und biserale Trennschärfe

Die Abbildung 3.4 zeigt die Korrelation von Itemschwierigkeit und Trennschärfe aller 45 Items. Der farbig hinterlegte Bereich beinhaltet die Items mit sowohl adäquater Itemschwierigkeit wie auch adäquater Trennschärfe. Insgesamt

28 der 45 Items wiesen eine adäquate Trennschärfe und Itemschwierigkeit auf. Die Items, welche nicht adäquat in Trennschärfe oder Itemschwierigkeit waren kann man weiter unterteilen in 8 Item, die zu leicht waren. Von diesen 8 Items zeigte ein 1 Item eine adäquate Trennschärfe. 9 Items mit adäquater Itemschwierigkeit zeigten keine Trennschärfe. In Tabelle 3.12 ist die Verteilung der Itemschwierigkeiten und in Tabelle 3.13 die Verteilung der Trennschärfen dargestellt.

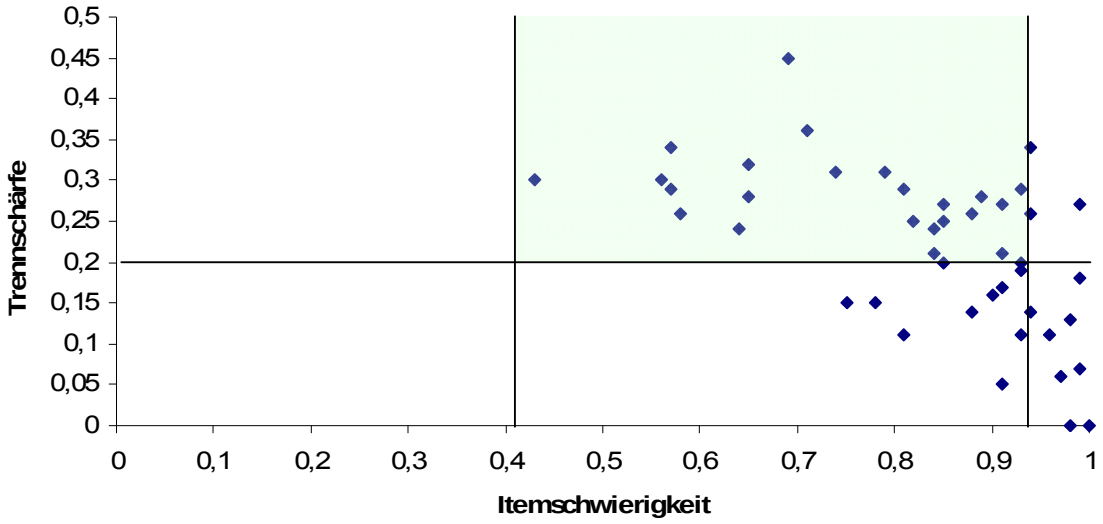

Abb. 3.4 Korrelation: Itemschwierigkeit – Trennschärfe
adäquate Trennschärfe ≥ 0,2, adäquate Itemschwierigkeit 0,41 - 0,94

Tab. 3.12 Verteilung der Itemschwierigkeiten

Beurteilung	p Wert	[n]	%
sehr schweres Item	≤ 20	0	0
schweres Item	21-40	0	0
adäquate Schwierigkeit	41-94	37	82
sehr leichtes Item	≥ 95	8	18

[n] Anzahl der Items ; % prozentuale Häufigkeit der Items

Tab. 3.13 Verteilung der Trennschärfe

Beurteilung	p Wert	[n]	%
negative Trennschärfe	< -0.09	0	0
ohne Trennschärfe	-0.09 - 0.09	5	11
schwache Trennschärfe	0.10 - 0.19	12	27
adäquate Trennschärfe	≥ 0.20	28	62

[n] Anzahl der Items ; % prozentuale Häufigkeit der Items

3.5 Objektivität

Als ein Maß für die Objektivität wurden die Bewertungen der Teststudierenden herangezogen. In Tabelle 3.14 wird die Differenz zum angestrebten Testwert bezogen auf die jeweilige Prüfungsstation und insgesamt angegeben. Spalte zwei zeigt die Punktzahl, auf welche die Teststudierenden trainiert wurden. Die Zeile „Punkte-Differenz“ gibt die Absolute Punktwertdifferenz als Betrag an. Zeile „Studierende-Differenz“ gibt an, wie viele Studierende (von insgesamt 2x8 = 16 Studierenden) abweichend von dem Trainingsergebnis bewertet wurden. Die Spalte „Gesamtdifferenz“ gibt die über alle Prüfungen ermittelte Gesamtdifferenz als Betrag, die Spalte „Insg. Bewertung“ als Absolutwert an.

Tab. 3.14 Bewertung der Teststudierenden

Test- Student	trainierte Pkt.	Parcours	BLS	Airway	Rhythmus	Defi	Trauma	Gesamt-diffenrenz	Insg. Bewertung
1	**6**	**1**	0	1	0	1	0	2	2
1	**6**	**2**	0	3	0	1	0	4	4
2	**9**	**1**	0	-1	0	0	0	1	-1
2	**9**	**2**	0	0	-1	0	-1	2	-2
3	**9**	**1**	0	0	0	-1	-1	2	-2
3	**9**	**2**	-1	0	0	0	-1	2	-2
4	**9**	**1**	0	-1	0	0	-1	2	-2
4	**9**	**2**	0	0	0	0	-1	1	-1
5	**6**	**1**	0	1	0	0	-1	2	0
5	**6**	**2**	1	0	-2	0	1	4	0
6	**6**	**1**	1	0	0	1	2	4	4
6	**6**	**2**	1	0	0	2	1	4	4
7	**9**	**1**	0	0	0	0	0	0	0
7	**9**	**2**	-1	0	-1	0	0	2	-2
8	**6**	**1**	1	3	-1	0	1	6	5
8	**6**	**2**	-1	0	0	1	1	3	1
Pkt-Differenz			7	10	5	7	12	41	8
Studierende Differenz			7	6	4	6	11	34	13

Von den Teststudierenden trainierte Punktzahl und deren Abweichung an den einzelnen Prüfungsstationen in jeweils zwei Durchgängen, sowie die Gesamtdifferenz als absolute Punktabweichung und die Insgesamt-Bewertung. Die Punktedifferenz an den jeweiligen Stationen wird angegeben sowie die Anzahl der Teststudierenden, die nicht ihre trainierten Punkte erreichten.

4. DISKUSSION

4.1 Gesamtergebnis

Die Gesamtergebnisse (Tab. 3.1) der OSCE-Prüfung vom Sommersemester 2005 ergeben, dass die Studierenden nach Absolvierung des Moduls Notfall- und Intensivmedizin die an sie gestellten Anforderungen erfüllen konnten. Nur ein Kandidat erreichte in der OSCE die Bestehensgrenze nicht.

Im Verlauf des Prüfungstages zeigten sich Verbesserungen bei der Globalbewertung bei Verschlechterung der Checklistenbewertung (Tab. 3.2). Da die Prüfung über den ganzen Tag stattfand, konnte nicht verhindert werden, dass Kandidaten, die früh geprüft wurden, ihren Kommilitonen, die die Prüfung noch vor sich hatten, Informationen über Stationen, Fragen und Abläufe geben. Diese Informationen scheinen den Prüflingen aber keinen deutlichen Vorteil zu erbringen, weil bei einer OSCE-Prüfung die Wiedergabe von Fakten allein nicht ausreicht, um ein gutes Ergebnis zu erzielen. Da die Prüfung parallel in zwei Prüfungsparcours durchgeführt wurde, konnte verhindert werden, dass ein zweiter Prüfungstag eingelegt werden musste und somit die Weitergabe von Informationen bis auf einen gewissen Anteil minimieren. Rutala et al. zeigten schon 1991, dass die evtl. stattfindende Weitergabe von Prüfungsinformationen innerhalb eines OSCE-Prüfungstages nicht zu signifikanten Veränderungen der Ergebnisse führt (Rutala et al. 1991). Die leichten Ergebnisschwankungen innerhalb des Prüfungstages könnten durch eine Änderung des Bewertungsmaßstabes der Prüfer zustande gekommen sein.

4.2 Videoparcours

Ein Vergleich der Prüfungsergebnisse im Bereich der Checklistenitems zwischen den beiden Prüfungsparcours A und B, von denen nur A anschließend in das Videorating einging, zeigt bei 15 Items einen signifikanten Unterschied in der Bewertung. Bei 11 dieser Items wurde im Parcours A häufiger das Item als nicht korrekt erfüllt bewertet. Nur bei vier Items bewerteten die Prüfer in Parcours A besser als in Parcours B. Eine Erklärung hierfür könnte sein, dass alle Prüfer

darüber informiert waren, dass Videoaufzeichnungen im Parcours A für anschließende Untersuchungen durchgeführt wurden, und somit eventuell genauer bzw. strenger bewerteten.

4.3 Objektivität

Mit ihrer Grundstruktur stellt die OSCE ein objektives Prüfungsinstrument dar. Durch kurze Prüfer-Prüfling Kontakte, den Einsatz von Checklisten mit definierten Lösungen und die Möglichkeit, viele Studierende mit denselben Aufgaben konfrontieren zu können, konnten bereits Harden et al. (Harden et al. 1975) die Objektivität dieser Prüfungsform zeigen.

Die durchgeführte Untersuchung mit den Teststudierenden bezüglich der Objektivität zeigt bei dieser Methode nur eine eingeschränkte Objektivität der Prüfung. Die Teststudierenden, die auf eine Bewertung mit drei Fehlern trainiert wurden, bekamen tendenziell eine etwas bessere Punktzahl. Nur in einigen Fällen wurden sie schlechter bewertet. Auffällig war, dass nur ein Studierender aus der Gruppe, die auf neun Punkte trainiert wurde, am Ende in einem Prüfungsparcours die volle Punktzahl erhielt. Bei allen Testkandidaten wurde nach jedem Parcours eine Befragung durchgeführt. Studierende, die schlechter bewertet wurden als ihre trainierte Punktzahl, gaben an, dass die trainierten Abläufe in der Prüfung nicht mehr korrekt durchgeführt oder Teile vergessen wurden. Kandidaten, die besser bewertet wurden, gaben an, Hilfestellungen von den Prüfern erhalten zu haben.

Somit wurden die Teststudierenden zum Großteil kongruent zu ihrer Prüfungsleistung bewertet. Allerdings zeigte diese Untersuchung, dass von den Prüfern während der Prüfung viele Hinweise gegeben wurden. Die Befragungen zeigten dabei, dass keineswegs jeder Teststudierende gleich viel Hilfestellung bekam. Ob das Maß der Hilfestellung abhängig von Sympathie ist, kann anhand dieser sehr kleinen Untersuchung nicht geklärt werden. Jefferies et al. konnten in einer ebenfalls sehr kleinen Studie zeigen, dass Vertrautheit keinen Einfluss auf die Beurteilung in einem OSCE hat (Jefferies et al. 2007). Burchard et al. beobachteten, dass vor allem die Bewertung von weniger kompetenten Kandidaten in einem OSCE den Prüfern Probleme bereitet. Deshalb fordern sie zwei Prüfer für eine OSCE-Station, um gerade schlechtere Performances in einer OSCE korrekt zu bewerten (Burchard et al. 1995).

Um exaktere Angaben bezüglich der Objektivität dieser Prüfung machen zu können, hätte man eine größere Anzahl an Teststudenten einsetzen müssen. Auch hätte das Training intensiviert und der Umgang mit Hilfestellungen vom Prüfer abgesprochen werden müssen.

4.4 Validität

Für die durchgeführte OSCE- Prüfung wurde im Vorfeld ein Blueprint erstellt, der sicherstellt, dass die Ziele des Curriculums in der Prüfung repräsentiert sind. Newble empfiehlt die Erstellung eines Blueprints, um eine hohe Inhaltvalidität zu gewährleisten (Newble 2004). Als ein weiteres notwendiges Kriterium für die Validität ist die Reliabilität zu nennen, auf die im Weiteren ausführlich eingegangen wird.

4.5 Reliabilität

4.5.1 Reliabilitätsformen

Der Reliabilitätsbegriff wird im Zusammenhang mit Objective Structured Clinical Examinations in der Literatur häufig angewendet (Chenot und Ehrhardt 2003; Cohen R. et al. 1990; Newble et al. 1981; Petrusa et al. 1990; Roberts und Norman 1990; Walters et al. 2005). Dabei unterscheiden sich die Reliabilitätsmaße in psychometrischen Tests von den Beurteilerreliabilitäten. In psychometrischen Tests sind drei Reliabilitätsbegriffe gebräuchlich: die RetestReliabiliät, die Split-half-Reliabilität und die „Interne Konsistenz“. Um die Retest-Reliabilität zu bestimmen, wird ein Test zu zwei Zeitpunkten von einem Probanden durchgeführt. Dabei wird hauptsächlich die zeitliche Stabilität der Testwerte erfasst. Split-half-Reliabilität und „Interne Konsistenz“ quantifizieren die Zuverlässigkeit eines Tests zu einem bestimmten Zeitpunkt. Bei der Split-half-Reliabilität wird der Test in zwei Hälften geteilt, um die Korrelation zu bestimmen. Die „Interne Konsistenz“ wird

ermittelt, indem der Zusammenhang zwischen den einzelnen Items mit der Gesamtheit der übrigen Items bestimmt wird (Lienert et al. 1998). Die Interraterreliabilität bezeichnet eine Korrelation der Messwerte bei Beurteilung durch zwei oder mehr Rater. Bei der Bewertung durch mehrere Rater wird bestimmt, wie präzise diese Rater die Ausprägung eines Items erkennen und nicht, ob alle Items eines Tests dieselbe Merkmalsdimension erfassen (Wirtz und Caspar 2002).

Die Vielseitigkeit des Reliabilitätsbegriffs erschwert einen Vergleich der vorliegenden Arbeit mit der Literatur. Eine weitere Problematik zeigt sich, wenn man den Aufbau der verschieden OSCEs in der Literatur vergleicht. Es gibt keine Standard-OSCE-Prüfung. Sowohl in Form wie auch im Ablauf zeigen sich deutliche Differenzen.

4.5.2 Reliabilität der OSCE-Prüfung

Allgemein wird in der Literatur ein Reliabilitätskoeffizient von $\kappa > 0,81$ für eine sehr gute Übereinstimmung angegeben. Ein κ zwischen 0,61 und 0,8 wäre ein Parameter für eine gute Reliabilität. Liegt der κ-Koeffizient im Bereich von 0,41-0,6, kann dies als akzeptable Übereinstimmung gewertet werden. Ob dieser Bereich noch akzeptabel ist, muss je nach Art des Tests abgewägt werden (Brennan und Silman 1992). Es gibt Autoren, die einen Reliabilitätskoeffizienten $\kappa > 0,5$ als zufriedenstellend einstufen (Maercker et al. 2004) wie auch Autoren , die deutlich strengere Maßstäbe wählen und erst ab einem $\kappa > 0,7$ von zufriedenstellender Übereinstimmung ausgehen (Bakeman und Gottman 1986). Für lizenzierende Examen fordert Downing sogar einen Reliabilitätkoeffizienten $\kappa > 0,9$ (Downing 2004). Für Prüfungen, die am Ende eines Moduls durchgeführt werden, kann auch ein niedrigerer κ Wert als akzeptabel gewertet werden.

Insgesamt besteht für die durchgeführte OSCE bei Betrachtung der Gesamtprüfung mittels Checklistenbewertung eine gute Reliabilität (Brennan und Silman 1992). Obwohl Newble und Swanson eine OSCE Prüfung mit vierstündiger Dauer fordern (Newble und Swanson 1988) und in der durchgeführten OSCE die Prüfdauer nur 25 Minuten pro Prüfling betrug, hat diese Prüfung mit einem Median vom $\kappa = 0,73$ für eine Modulabschlussprüfung eine gute Reliabilität.

Nikendei und Jünger beobachteten bei einem interdisziplinären Innere Medizin OSCE mit 12 Stationen einen Reliabiliätskoeffizienten Crohnbach alpha von 0,748 (Nikendei und Jünger 2006). Crohnbachs alpha wird angewendet, um die Split-half-Reliabilität zu berechnen. Mit diesem Koeffizienten besteht die Gefahr die Reliabilität, zu überschätzen (Schmidt et al. 2003). Roberts und Norman zeigten schon 1990, dass die Reliabilität zwischen den Stationen berechnet mit Crohnbachs alpha (α= 0,198) deutlich geringer ist, als die Reliabilität der einzelnen Stationen. Dabei wurden einige Stationen mit zwei Prüfern besetzt und somit eine Interraterreliabilität mittels Intraklassenkoeffizient (ICC=0,80-0,99) berechnet. Ebenfalls wurde die Reliabilität von Stationen beobachtet, die von einigen Studenten zweimal durchlaufen wurden. Diese sogenannte Test-Retest-Reliabilität lag zwischen 0,66-0,86 (Roberts und Norman 1990). Chenot et al. führten eine Untersuchung in einem Basisfähigkeiten OSCE durch und berechneten eine Reliabilitätskoeffizienten Kappa von 0,66 (Chenot et al. 2007). Dabei wurden 4 der 10 Stationen sowohl von ärztlichen wie auch studentischen Prüfern bewertet und dann der Reliabilitätskoeffizient berechnet.

Die Reliabilität für die einzelnen Stationen schwankt von guten Ergebnissen im Bereich der „BLS“-Prüfung κ=0,63, der „Rhythmus“-Station κ=0,63 und der „Airway“-Prüfung κ=0,65, zu moderaten Kappa-Koeffizienten an der Prüfungs-Station „Defi“ κ=0,53 und der „Trauma“-Station mit κ =0,44.

In einer von Vivekanada-Schmidt et. al. durchgeführten videobasierten OSCE-Analyse sind zwei Stationen mittels Videoanalyse auf die Reliabilität untersucht worden. Die Reliabilität für die Checklistenbewertung lag bei κ= 0,43 und κ=0,51 und zeigt damit eine etwas geringere Reliabilität als die vorliegende Untersuchung (Vivekananda-Schmidt et al. 2007). O’Connor und Mc Graw untersuchten in einer 18 Stationen OSCE zwei Stationen mit jeweils doppelter Prüferbesetzung auf Interraterreliabilität. Bei den Prüfern handelte es sich um Fachärzte aus den Bereichten Chirurgie, Innere Medizin und Allgemeinmedizin sowie Assistenzärzte aus der Chirurgie und der Inneren Medizin. Alle bekamen mindestens 30 Minuten vor Beginn der Prüfung ein Handout mit Erläuterungen zur Checkliste und wurden in die Prüfung eingewiesen. Kappa wurde mit 0,65 für die Naht-Station bzw. 0,71 für die Intubations-Station angegeben. Die Reliabilität dieser beiden Prüfungstationen liegt somit im zufriedenstellenden Bereich und

zeigt ähnliche bzw. leicht höhere Werte als die Prüfung in der vorliegenden Arbeit (O'Connor und McGraw 1997).

Insgesamt betrachtet fällt auf, dass die Gesamtprüfung eine gute Übereinstimmung aufweist, während bei den einzelnen Stationen eine niedrigere Reliabilität herrscht. Dies bestätigt einerseits die Forderung nach OSCE Prüfungen mit ausreichender Dauer und Stationenanzahl (Newble und Swanson 1988), andererseits wird deutlich, dass OSCE-Ergebnisse nur als Gesamtergebnis und nicht für Einzelstationen betrachtet werden sollten. Dies ist insbesondere für den Lerneffekt der Prüfung und die Zufriedenheit der Studenten limitierend, da neben der ausbleibenden Rückmeldung direkt nach einer Station nun auch nur die Einsicht in die Gesamtleistung der OSCE gewährt werden kann. Allen et. al. zeigten, dass ein eineinhalbminütiges Feedback am Ende einer jeden OSCE Station von den Prüfungskandidaten als wenig störend und vor allem hilfreich empfunden wurde (Allen et al. 1998). Ein solches Feedback führt auch zu einer Verbesserung der studentischen Fertigkeiten (Hodder et al. 1989).

Seit einigen Jahren wurden bei guter Reliabilität zunehmend Globalbeurteilungen in OSCE-Prüfungen verwendet (Hodges und McIlroy 2003). In der durchgeführten OSCE-Prüfung liegt der Reliabilitätskoeffizient Kappa für die Globalbewertung der einzelnen Stationen mit Werten zwischen κ=0,33-0,43 nicht in einem zufriedenstellenden Bereich. Die Fachärzte unter den Videorater zeigten durchgehend einen niedrigeren Reliabilitätskoeffizienten Kappa als die Studenten und Assistenten. Regehr et al. zeigen, dass eine Globalbeurteilung in einer OSCE eine höhere Reliabilität zeigt als die Checklistenbewertung, wenn die Beurteilung von erfahrenen Prüfern durchgeführt wird (Regehr et al. 1998). McIlroy et al. konnten dagegen zeigen, dass die Reliabilität bei Globalbeurteilung etwas niedriger ist. Des Weiteren ändert sich das Verhalten der Prüfungskandidaten, wenn sie vorher erfahren, ob sie anhand einer Checkliste oder mittels Globalbeurteilung bewertet werden. Die Globalbewertung mit Hilfe einer Likert-Skala wird vor allem für die Bewertung von kommunikativen Fähigkeiten, Verhalten und zur Beurteilung von Problemlösungsstrategien eingesetzt. Prüflinge, die eine Checklistenbewertung erwarten, zeigen ein einstudiertes Ablaufschema, während Kandidaten, die eine Globalbeurteilung erwarten mehr Wert auf Verhalten und Kommunikation legen, als auf ein standardisiertes Schema (McIlroy

et al. 2002). In der vorliegenden OSCE-Prüfung war die Reliabilität der Globalbeurteilung niedriger als die Reliabilität der Checklistenbewertung. Dieses Ergebnis kann darauf zurückgeführt werden, dass an allen Prüfungsstationen vermehrt Wert auf ein korrektes Vorgehen gelegt wurde. In der durchgeführten OSCE Prüfung gab es keine Anamnesestationen, weder reale Patienten noch Simulationspatienten wurden eingesetzt, was dazu führte, dass von den Prüflingen nur wenig kommunikative Fähigkeit gefordert wurde. Allerdings sollte auch beachtet werden, dass es sich bei der Globalbewertung in dieser Modulabschlussprüfung um eine einzelne Globalbewertung pro Station ohne jegliche Unterpunkte handelte. Beim Einsatz von Globalbeurteilungen hat das Training der Beobachter deutlich mehr Relevanz, damit diese Gleiches möglichst gleich bewerten (Krebs 1999). Mangelnde Reliabilität einer Prüfung kann dadurch verursacht werden, dass sich Rater nicht darüber einig sind, welches Merkmal beurteilt werden soll. Durch genaue Formulierungen der Items innerhalb der Checkliste wurde versucht, den Ratern eine präzise Beschreibung der Anforderung darzustellen. Eine weitere Ursache für unzureichende Reliabilität eines Tests ist die Tatsache, dass Beurteiler die Ausprägung von Merkmalen unterschiedlich beurteilen. Der Schwellenwert, ab dem ein Merkmal oder Verhalten als „vorhanden“ gewertet wird, unterscheidet sich bei den Beurteilern (Wirtz und Caspar 2002). In der vorliegenden Arbeit wurden alle Videorater ca. 30 min eingewiesen. Dabei wurde vermehrt Wert auf das korrekte Ausfüllen der Prüfungsbögen gelegt. Die Prüfer in der Originalprüfung erhielten eine ca. 1,5 h Einweisung mit Begehung der Stationen und Probedurchlauf jeder Station. Retrospektiv kann angenommen werden, dass die Einweisung der Videorater einen anderen Schwerpunkt inne hatte und somit die Bewertungsmaßstäbe der Videorater stark divergierten.

4.6 Vergleich studentischer – ärztlicher Prüfer

In der vorliegenden Arbeit haben studentische wie auch ärztliche Prüfer im Videorating eine adäquate Bewertung der Prüflinge durchgeführt. Dies konnte über einen zufriedenstellenden Zusammenhang mittels Spearman r_s dargestellt werden. Spearman r_s bei den studentischen Ratern liegt zwischen 0,57-0,87 bei

den ärztlichen Ratern zwischen r_s 0,43-0,83 für die Checklistenbewertung. Die maximale Punktabweichung einer Station lag bei den Studenten bei 0,51 und bei den Ärzten bei 0,48 Punkten. Wobei die Studenten weniger Punkte vergeben haben als die Ärzte. Dies wird auch in Abbildung 3.1 anhand der Punktedifferenz deutlich. In dieser Abbildung zeigt sich des Weiteren, dass nur ein sehr geringer Anteil an Prüflingen die identische Punktzahl im Videorating wie in der Originalprüfung erhalten hat. Bei der Globalbewertung lag der Spearman r_s bei den studentischen Ratern zwischen 0,57-0,75 und bei den ärztlichen Ratern bei 0,48-0,73. Insgesamt haben die Studenten die Prüflinge bei der Globalbeurteilung etwas besser bewertet als in der Originalprüfung und die Ärzte etwas schlechter. Die maximale Notendifferenz lag bei den Studenten bei 0,21 und bei den Ärzten bei 0,66. Die Übereinstimmung mit der Originalprüfung in Checklisten- wie auch Globalbeurteilung war bei den Studenten größer als bei den Ärzten mit Ausnahme der Airway-Station.

Chenot et al. zeigten, dass studentische Tutoren reliabel (Kappa 0,41- 0,66) bewerten können. Die maximale Notendifferenz wird mit 0,28 angegeben. Die studentischen Tutoren wurden zuvor intensiv in die Studentenausbildung in Kleingruppenunterricht als Hilfskräfte mit eingebunden (Chenot et al. 2007). Ebenfalls gute Übereinstimmungen von Studenten und Ärzten konnte in einer zahnmedizinischen Untersuchung gezeigt werden (Ogden et al. 2000). Von 125 Kandidaten wurden nur 9 mit mehr als einem Punkt Unterschied bewertet. Auch diese studentischen Prüfer nahmen vorher als Tutoren an der Studentenausbildung teil. Bei beiden Studien muss beachtet werden, dass evtl. Absprachen zwischen studentischen und ärztlichen Prüfern stattgefunden haben, da sich die Prüfer gemeinsam in einem Raum befanden.

Studenten konnten in der vorliegenden Untersuchung die OSCE-Teilnehmer in den Videoaufzeichnungen reliabel bewerten. Diese Ergebnisse beziehen sich allerdings auf eine Videobeurteilung und können nicht unkritisch auf eine reale Prüfungssituation übertragen werden. Die ausgewählten Studenten der Videoanalyse in der vorliegenden Arbeit nahmen nicht als Tutoren an der Ausbildung der Studenten teil, sind aber im Bereich Notfallmedizin interessiert und engagiert

4.7 Itemschwierigkeit und Trennschärfe

In der vorliegenden Arbeit zeigen mehr als 80% aller Items eine adäquate Itemschwierigkeit. Bei den restlichen Items handelt es sich um leichte Items. Im Sinne der klassischen Testtheorie ist es unter dem Aspekt der guten Differenzierung überflüssig, „leichte“ Fragen zu stellen (Möltner et al. 2006). Es kann argumentiert werden, dass leichte Items sogenannte „Eisbrecheritems“ sind, die den Kandidaten den Einstieg in die Prüfung erleichtern. In der durchgeführten OSCE-Prüfung stehen allerdings nur drei der acht leichten Items am Beginn der Prüfungsstation und die „Eisbrecher“-Funktion ist durchaus fraglich, da die Prüflinge die Items der Checkliste nicht kannten und außerdem kein Feedback am Ende der Prüfung erhielten Die Aufgabenschwierigkeit stellt keine eigentliche Eigenschaft der Aufgabe dar, sondern ist immer im Bezug zur jeweiligen geprüften Stichprobe zu sehen. Somit ist es in einer kriteriumsorientierten Prüfung nicht negativ, wenn die einzelnen Aufgaben relativ leicht sind, da diese als wichtige basale Kenntnisse im Vorfeld der Prüfung festgelegt wurden. Der überwiegende Teil der Prüfungsaufgaben zeigt eine adäquate Trennschärfe. Nur ein geringer Anteil der Items weist keine Trennschärfe auf. Bei den Items ohne Trennschärfe handelt es sich ausschließlich um Aufgaben mit hoher Itemschwierigkeit. Somit hängt die mangelnde Trennschärfe mit zu „leichten“ Aufgaben zusammen und beruht nicht auf schlecht formulierten Items. Die fehlende Trennschärfe von Aufgaben führt nicht zu einer Minderung der Reliabilität. Wie oben erwähnt, dienen Prüfungen dazu, grundlegende Fertigkeiten zu prüfen und nicht dazu, Abstufungen zwischen „guten“ und „schlechten“ Studierenden aufzuzeigen. Da durch eine gute Lehre Basiswissen ausreichend vermittelt wird, werden die meisten Prüfungsaufgaben durch einen Großteil der Studierenden erfolgreich erledigt. Diese Aufgaben zeigen dann eine geringe Trennschärfe und sind zu leicht. Trotzdem sollte nicht versucht werden, die Gütemaße um jeden Preis zu steigern und somit eine unangemessene Prüfung mit sogenannten „Kolibri-Fragen“ zu konstruieren (Möltner et al. 2006).

Auch Nikendei und Jünger empfehlen für OSCEs neben Aufgaben mit adäquater Trennschärfe, die kompetente von inkompetenten Kandidaten trennen, auch Aufgaben, die Basiswissen abfordern und somit eher niedrige Trennschärfe aufweisen (Nikendei und Jünger 2006).

5. Zusammenfassung

Mit Einführung der neuen Approbationsordnung für Ärzte erfolgten ab dem Sommersemester 2004 durch die Universitäten zahlreiche Umstrukturierungen und Neuerungen. Für den Querschnittsbereich Notfall- und Intensivmedizin wurde in Göttingen eine praktische Prüfung im Sinne einer Objektive Structured Clinical Examination (OSCE) eingeführt. Zunächst erfolgte die Planung einer validen Prüfung mittels Blueprint mit den vorhanden Ressourcen. Prüfungsziele wurden festgelegt und Prüfungsbögen entworfen. Anschließend wurde die OSCE auf die beiden weiteren Qualitätskriterien einer Prüfung, Reliabilität und Objektivität, untersucht und eine Itemanalyse durchgeführt. Aufgrund von limitierten Ressourcen im personellen, zeitlichen und materiellen Bereich konnte nur eine OSCE-Prüfung mit fünf Stationen durchgeführt werden.

Die OSCE-Prüfung im Sommersemester 2005 wurde mittels eines Videoratings auf Reliabilität untersucht. Teilnehmer waren 91 Studenten aus dem 5. klinischen Semester. Am Videorating nahmen neben ärztlichen Ratern auch studentische Rater teil. Die Reliabilität wurde mit Hilfe des Reliabilitätskoeffizienten Kappa berechnet. Eine Global- und eine Checklistenbewertung wurden miteinander verglichen. Die Untersuchung zur Objektivität fand bereits ein Semester zuvor mit acht Teststudierenden statt.

Die durchgeführte OSCE-Prüfung erfüllte mit ihrer nur 25 minütigen Dauer nicht die geforderte Länge einer reliablen OSCE-Prüfung. In der Untersuchung zeigte die OSCE mit einem κ=0,73 im Bereich der Checklistenbewertung eine gute Reliabilität. Schwächen zeigten sich in einzelnen Prüfungsstationen, wie der Defi- oder der Traumastation. Eine akzeptable Reliabilität konnte allerdings immer gewährleistet werden. Die Globalbewertung zeigte eine insgesamt nicht zufriedenstellende Reliabilität κ=0,33-0,43. Allerdings ist in dieser Prüfung ausschließlich eine Globalbewertung pro Station erfolgt und nicht eine Globalbewertung für jedes Item. Der Vergleich studentischer Rater (r_s 0,57-0,87) mit den ärztlichen Ratern (r_s 0,43-0,83) zeigt zufriedenstellende bis gute Übereinstimmung mit der Originalprüfung für die Checklistenbewertung. Insgesamt wurden von den studentischen Prüfern weniger Punkte für die Leistungen der Prüflinge in der Checklistenbewertung vergeben als von den ärztlichen Prüfern. Bei der Globalbewertung dagegen haben die studentischen

Rater bessere Noten vergeben als die ärztlichen Rater. Die Übereinstimmung mit der Originalprüfung war bei der Globalbeurteilung etwas schlechter (r_s 0,48-0,75). Ingesamt war die Übereinstimung mit der Originalprüfung bis auf die Airway-Station immer besser bei den studentischen Ratern. Die Untersuchung zur Objektivität der Prüfung zeigte Verbesserungspotential in diesem Bereich. Die Itemanalyse ergab für 28 Items eine adäquate Itemschwierigkeit sowie adäquate Trennschärfe.

Die OSCE- Prüfung im Bereich Notfallmedizin in Göttingen ist eine reliable Prüfungsform. Im verwendeten Prüfungsbogen ist die Checklistenbewertung der Globalbewertung im Bereich Reliabilität überlegen. Studenten, die im Videorating eingesetzt wurden, konnten die Prüfung mit guter Übereinstimmung zur Originalprüfung bewerten.

6. ANHANG

6.1 Der Prüfungsparcours

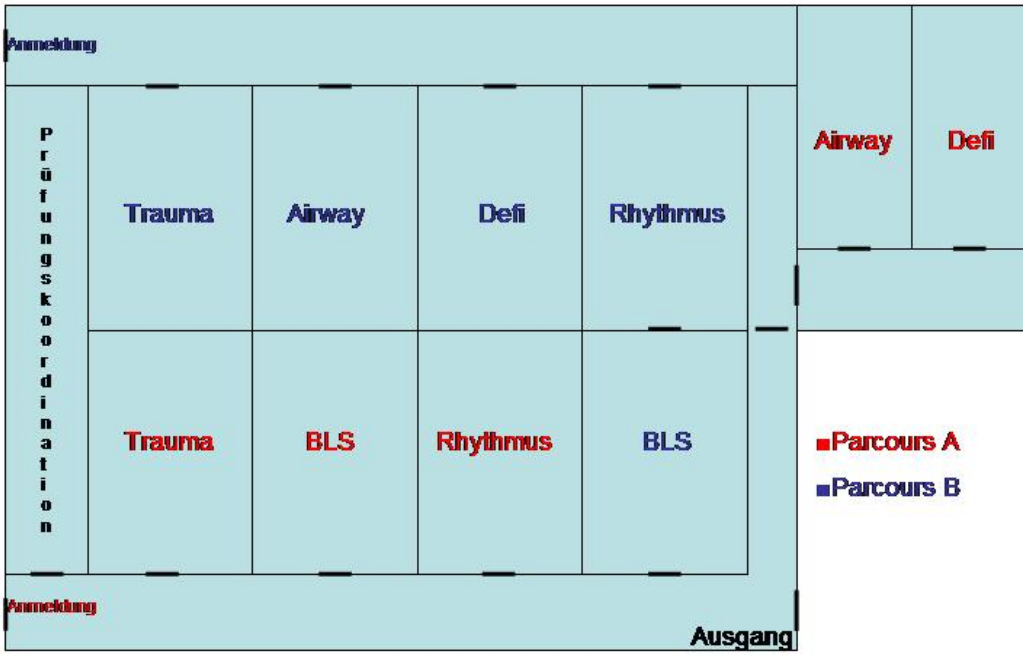

Abb. 6.1 Aufbau des Prüfungsparcours

6.2 Zeitplan OSCE-Prüfung

OSCE-Prüfung M6.2

Prüfungsort: Lehr- und Simulationszentrum, UBFT B4 01 Raum 877
Anmeldung: 15min vor Prüfungsbeginn
mit Personal-/Studentenausweis und Evaluationsbogen

Prüfungsdauer: max. 90min

Parcours	**A**	**B**
Zeit	**Gruppe**	**Gruppe**
08.00	1	2
08.30	3	4
09.00	5	6
09.30	7	8
10.00	9	10
10.30	11	12
11.00	13	14

Parcours	**A**	**B**
Zeit	**Gruppe**	**Gruppe**
12.00	15	16
12.30	17	18
13.00	19	20
13.30	21	22
14.00	23	24
14.30	25	26
15.00	27	28

7. Literaturverzeichnis

Ali J, Cohen R, Adam R, Gana TJ, Pierre I, Ali E, Bedaysie H, West U, Winn J (1996a): Attrition of cognitive and trauma management skills after the Advanced Trauma Life Support (ATLS) course. J Trauma 40 (6), 860-866

Ali J, Cohen R, Adam R, Gana TJ, Pierre I, Bedaysie H, Ali E, West U, Winn J (1996b): Teaching effectiveness of the advanced trauma life support program as demonstrated by an objective structured clinical examination for practicing physicians. World J Surg 20 (8), 1121-1125

Ali J, Adam R, Williams JI, Bedaysie H, Pierre I, Josa D, Winn J (2002): Teaching effectiveness of the trauma evaluation and management module for senior medical students. J Trauma 52 (5), 847-851

Allen R, Heard J, Savidge M, Bittergle J, Cantrell M, Huffmaster T (1998): Surveying Students' Attitudes During the OSCE. Adv Health Sci Educ Theory Pract 3 (3), 197-206

Bakeman R, Gottman JM: Observing interaction: an introduction to sequential analysis. Cambridge University Press, Cambridge 1986

Barman A (2005): Critiques on the Objective Structured Clinical Examination. Ann Acad Med Singapore 34 478-482

Barrows HS (1968): Simulated patients in medical teaching. Can Med Assoc J 98 (14), 674-676

Baubin M, Dirks B (2008): Ausbildungskonzepte des European Resuscitation Councils (ERC). Notfall & Rettungsmedizin 11 276-278

Beckers S, Bickenbach J, Fries M, Hoffmann N, Classen-Linke IK, B., Wainwright U, Kuhlen R, Rossaint R (2004): "Meet the AIX-PERTs." Der notfallmedizinische Start in den Modellstudiengang Humanmedizin der Universität Aachen. Anaesthesist 53 561-569

Bloch R, Hofer D, Krebs R, Schläppi P, Weis S, Westkämper R: Kompetent prüfen. Handbuch zur Planung, Durchführung und Auswertung von Facharztprüfungen. Universität Bern Institut für Aus-, Weiter- und Fortbildung, Verbindung der Schweizer Ärzte (FMH) und Österreichischen Ärztekammer (ÖÄK), Bern 1999

Bortz J, Lienert GA: Kurzgefaßte Statistik für die klinische Forschung: ein praktischer Leitfaden für die Analyse kleiner Stichproben. Springer, Berlin 1998

Bortz J, Döring N: Klassische Testtheorie; in: Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler.; hrsg. v. Bortz J, Döring N; Springer, Berlin 2002, 192-202

Brennan P, Silman A (1992): Statistical methods for assessing observer variability in clinical measures. BMJ 304 (6840), 1491-1494

Bundesministerium für Gesundheit (2002): Approbationsordnung für Ärzte, ausgegeben zu Bonn am 3. Juli 2002. Bundesgesetzblatt 44 2405-2435

Burchard KW, Rowland-Morin PA, Coe NP, Garb JL (1995): A surgery oral examination: interrater agreement and the influence of rater characteristics. Acad Med 70 (11), 1044-1046

Burdick WP, Ben-David MF, Swisher L, Becher J, Magee D, McNamara R, Zwanger M (1996): Reliability of performance-based clinical skill assessment of emergency medicine residents. Acad Emerg Med 3 (12), 1119-1123

Chenot JF, Ehrhardt M (2003): Objective structured clinical examination (OSCE) in der medizinischen Ausbildung: eine Alternative zur Klausur. Z Allg Med 79 437-442

Chenot JF, Fischer T, Simmenroth-Nayda A, Fassheber S, Hummers-Pradier E, Aut B, Kernbach-Wighton G, Emmert S, Küntzel H, Klockgether-Radke AP, Kochen MM (2004): Interdisziplinärer Pilot- OSCE - "Medizinische Basisfähigkeiten". Z Allg Med 80 503-506

Chenot JF, Simmenroth-Nayda A, Koch A, Fischer T, Scherer M, Emmert B, Stanske B, Kochen MM, Himmel W (2007): Can student tutors act as examiners in an objective structured clinical examination? Med Educ 41 (11), 1032-1038

Cohen J (1960): A coefficient of agreement for nominal scales. Educ Psychol Meas 20 37-46

Cohen J (1968): Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70 (4), 213

Cohen R, Reznick RK, Taylor BR, Provan J, Rothman A (1990): Reliability and validity of the objective structured clinical examination in assessing surgical residents. Am J Surg 160 (3), 302-305

Downing SM (2004): Reliability: on the reproducibility of assessment data. Med Educ 38 (9), 1006-1012

Dupras DM, Li JT (1995): Use of an objective structured clinical examination to determine clinical competence. Acad Med 70 (11), 1029-1034

European Resuscitation Council (2000): International Guidelines 2000 for CPR and ECC: A Consensus on Science. Resuscitation 46 (1-3), 1-447

Georg-August-Universität Göttingen: Studienordnung des Studiengangs Humanmedizin an der Georg- August- Universität Göttingen 2004; Georg-August-Universität Göttingen; Göttingen 2004.

Harden RM, Stevenson M, Downie WW, Wilson GM (1975): Assessment of clinical competence using objective structured examination. BMJ 1975;1 447-451

Harden RM, Gleeson FA (1979): Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 13 (1), 41-54

Heckmann JG, Dutsch M, Rauch C, Lang C, Weih M, Schwab S (2008): Effects of peer-assisted training during the neurology clerkship: a randomized controlled study. Eur J Neurol 15 (12), 1365-1370

Hill D, Stalley P, Pennington D, Besser M, McCarthy W (1997): Competency-based learning in traumatology. Am J Surg 173 (2), 136-140

Hodder RV, Rivington RN, Calcutt LE, Hart IR (1989): The effectiveness of immediate feedback during the objective structured clinical examination. Med Educ 23 (2), 184-188

Hodges B, McIlroy JH (2003): Analytic global OSCE ratings are sensitive to level of training. Med Educ 37 (11), 1012-1016

Jefferies A, Simmons B, Regehr G (2007): The effect of candidate familiarity on examiner OSCE scores. Med Educ 41 (9), 888-891

Johnson G, Reynard K (1994): Assessment of an objective structured clinical examination (OSCE) for undergraduate students in accident and emergency medicine. J Accid Emerg Med 11 (4), 223-226

Junger J, Schafer S, Roth C, Schellberg D, Friedman Ben-David M, Nikendei C (2005): Effects of basic clinical skills training on objective structured clinical examination performance. Med Educ 39 (10), 1015-1020

Krebs R: Wie wird eine objektive strukturierte klinische Prüfung entwickelt? in: Kompetent prüfen. Handbuch zur Planung, Durchführung und Auswertung von Facharztprüfungen.; hrsg. v. Universität Bern Institut für Aus- W-, und Fortbildung, Verbindung der Schweizer Ärzte (FMH) und Österreichischen Ärztekammer (ÖÄK), Bern 1999, 141-178

Levine HG, McGuire CH, Leroy William Nattress J (1970): The Validity of Multiple Choice Achievement Tests as Measures of Competence in Medicine. Am Educ Res J 7 (1), 69-82

Li MS, Brasel KJ, Schultz D, Falimirski ME, Stafford RE, Somberg LB, Weigelt JA (2006): Effective retention of primary survey skills by medical students after participation in an expanded Trauma Evaluation and Management course. Am J Surg 191 (2), 276-280

Lienert GA, Raatz U, Lienert R: Testaufbau und Testanalyse. 6. Aufl., Studienausg; Beltz Psychologie Verlags Union, Weinheim 1998

Lunenfeld E, Weinreb B, Lavi Y, Amiel GE, Friedman M (1991): Assessment of emergency medicine: a comparison of an experimental objective structured clinical examination with a practical examination. Med Educ 25 (1), 38-44

Maercker A, Michael T, Fehm L, Becker ES, Margraf J (2004): Age of traumatisation as a predictor of post-traumatic stress disorder or major depression in young women. Br J Psychiatry 184 482-487

Marton F, Saljo R (1976): On qualitative differences in learning: II - outcomes as a function of the learner´s conception of the task. Br J Educ Psychol 46 115-127

Mau W, Kusak G (2005): [Implementation of the new Federal Medical Licensing Regulations for doctors in the interdisciplinary subject "Rehabilitation, physical medicine, naturopathic treatment" by the German medical faculties]. Rehabilitation (Stuttg) 44 (3), 129-133

McIlroy JH, Hodges B, McNaughton N, Regehr G (2002): The effect of candidates' perceptions of the evaluation method on reliability of checklist and global rating scores in an objective structured clinical examination. Acad Med 77 (7), 725-728

McLaughlin K, Gregor L, Jones A, Coderre S (2006): Can standardized patients replace physicians as OSCE examiners? BMC Med Educ 6 12

Miller GE (1990): The assessment of clinical skills/competence/performance. Acad Med 65 (9), 63-67

Möltner A, Schellenberg D, Jünger J (2006): Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild 23 (3),

Newble DI (1992): Assessing clinical competence at the undergraduate level. Med Educ 26 504-511

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

Newble DI, Entwistle NJ (1986): Learning styles and approaches: implications for medical education. Med Educ 20 162-175

Newble DI, Swanson DB (1988): Psychometric characteristics of the objective structured clinical examination. Med Educ 22 (4), 325-334

Newble DI, Baxter A, Elmslie RG (1979): A comparison of multiple-choice tests and free-response test in examinations of clinical competence. Med Educ 13 263-268

Newble DI, Hoare J, Sheldrake PF (1980): The selection and training of examiners for clinical examinations. Med Educ 14 (5), 345-349

Newble DI, Hoare J, Elmslie RG (1981): Validity and reliability of a new examination of the clinical competence of medical students. Med Educ 15 46-52

Nikendei C, Jünger J (2006): OSCE - praktische Tipps zur Implemetierung einer klinisch- praktischen Prüfung. GMS Z Med Ausbild 23 (3), Dok47

O'Connor HM, McGraw RC (1997): Clinical skills training: developing objective assessment instruments. Med Educ 31 (5), 359-363

Ogden GR, Green M, Ker JS (2000): The use of interprofessional peer examiners in an objective structured clinical examination: can dental students act as examiners? Br Dent J 189 (3), 160-164

Petrusa ER, Blackwell TA, Rogers LP, Saydjari C, Parcel S, Guckian JC (1987): An objective measure of clinical performance. Am J Med 83 (1), 34-42

Petrusa ER, Blackwell TA, Ainsworth MA (1990): Reliability and validity of an objective structured clinical examination for assessing the clinical performance of residents. Arch Intern Med 150 (3), 573-577

Regehr G, MacRae H, Reznick RK, Szalay D (1998): Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med 73 (9), 993-997

Roberts J, Norman G (1990): Reliability and learning from the objective structured clinical examination. Med Educ 24 (3), 219-223

Roth E: Sozialwissenschaftliche Methoden: Lehr- und Handbuch für Forschung und Praxis. Oldenbourg, München 1984

Rutala PJ, Witzke DB, Leko EO, Fulginiti JV, Taylor PJ (1991): Sharing of information by students in an objective structured clinical examination. Arch Intern Med 151 (3), 541-544

Scheffer S, Muehlinghaus I, Froehmel A, Ortwein H (2008): Assessing students' communication skills: validation of a global rating. Adv Health Sci Educ Theory Pract 13 (5), 583-592

Schmidt FL, Le H, Ilies R (2003): Beyond alpha: an empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. Psychol Methods 8 (2), 206-224

Schrauth M, Schmulius N, Zipfel S, Haarmeier T (2006): [Practical examinations for neurology. The Tuebingen model]. Nervenarzt 77 (12), 1464-1468

Schumacher J, Brähler E: Psychodiagnostische Testverfahren; in: Leitfaden psychosomatische Medizin und Psychotherapie; hrsg. v. Jansson PL, Joraschky P, Tress W; Deutscher Ärzteverlag, Köln 2006, 86-92

Schwarzkopf SR, Morfeld M, Gulich M, Lay W, Horn K, Mau W (2007): [Current teaching, learning and examination methods in medical education and potential applications in rehabilitative issues]. Rehabilitation (Stuttg) 46 (2), 64-73

Simpson MA: Medical education: A critical approach.; Butterworths, London 1972

Timmermann A, Roessler M, Barwing J, Blaschke S, Brauer A, Eich C, Hirn A, Klockgether-Radke A, Nickel E, Russo S, Kettler D, Saur P (2005): [New pathways in undergraduate medical education - first experiences with the cross section speciality emergency and intensive care medicine.]. Anasthesiol Intensivmed Notfallmed Schmerzther 40 (9), 536-543

van der Vleuten CPM (1996): The assessment of professional competence: Developments, research and practical implications. Adv Health Sci Educ 1 41-67

van der Vleuten CPM, Schuwirth LW (2005): Assessing professional competence: from methods to programmes. Med Educ 39 (3), 309-317

Verma M, Singh T (1993): Experiences with objective structured clinical examination (OSCE) as a tool for formative evaluation in pediatrics. Indian Pediatr 30 699-702

Vivekananda-Schmidt P, Lewis M, Coady D, Morley C, Kay L, Walker D, Hassell AB (2007): Exploring the use of videotaped objective structured clinical examination in the assessment of joint examination skills of medical students. Arthritis Rheum 57 (5), 869-876

Walters K, Osborn D, Raven P (2005): The development, validity and reliability of a multimodality objective structured clinical examination in psychiatry. Med Educ 39 (3), 292-298

Wass V, van der Vleuten C (2004): The long case. Med Educ 38 (11), 1176-1180

Weißer F, Dirks B, Georgieff M (2004): Objective Structured Clinical Examination (OSCE): Eine neue Prüfungsform in der notfallmedizinischen Ausbildung. Notfall & Rettungsmedizin 7 (4), 237-243

Wilkinson TJ, Fontaine S (2002): Patients' global ratings of student competence. Unreliable contamination or gold standard? Med Educ 36 (12), 1117-1121

Wirtz M, Caspar F: Beurteilerübereinstimmung und Beurteilerreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen. Hogrefe Verl. für Psychologie, Göttingen 2002

Ziegler JS, Wagner T (2008): Implementierung eines Gremiums zur Koordination der Lehrverbesserung: Frankfurter Ideenforum für Lehre und Unterricht. GMS Z Med Ausbild. 25 (1), Doc 52

Danksagung

.

Ich danke Herrn Prof. Dr. M. Quintel und meinem Doktorvater Herrn PD. Dr. A. Timmermann für die freundliche Überlassung des Themas der Dissertation.

Besonders bedanken möchte ich mich bei PD. Dr. Arnd Timmermann für die Betreuung meiner Arbeit und die vielen wichtigen Anregungen.

Des Weiteren möchte ich dem Team des Lehr- und Simulatonszentrums für die tatkräftige Unterstützung danken. Ebenso bedanke ich mich bei der Anästhesietechnik und der Abteilung für Medizinische Statistik für die Beratung.

Lebenslauf

Am 4. November 1980 wurde ich, Katrin Schwerdtfeger, als zweites Kind von Wolfgang und Ursula Schwerdtfeger, geb. Schmidt, in Seesen geboren.

Im Sommer 1987 wurde ich in die Grundschule in Seesen eingeschult. Ab dem Sommer 1991 besuchte ich für zwei Jahre die Orientierungsstufe in Seesen und wechselte danach auf das Jacobson-Gymnasium Seesen.

Nach meinem Abitur im Sommer 2000 begann ich mit dem Medizinstudium im Wintersemester 2000/01 an der Georg-August-Universität Göttingen. Im März 2003 absolvierte ich das Physikum und im März 2004 das 1.Staatsexamen.

Das praktische Jahr begann ich im Sommer 2006. Die ersten beiden Tertiale absolvierte ich am Klinikum Oldenburg in der Abteilung Innere Medizin und der Abteilung Anaesthesiologie, Intensivmedizin, Notfallmedizin und Schmerztherapie. Das Chirurgie Tertial verbrachte ich im Spital Laufenburg in der Schweiz. Im November 2007 absolvierte ich die 2. Ärztliche Prüfung in Göttingen.

Seit dem 01.01.2008 arbeite ich als Assistenzärztin im Zentrum Anaesthesiologie, Rettungs- und Intensivmedizin der Universität Göttingen.