Aus der Klinik für Mund-, Kiefer- und Gesichtschirurgie
(Prof. Dr. med. Dr. med. dent. H. Schliephake)
im Zentrum Zahn-, Mund- und Kieferheilkunde
der Medizinischen Fakultät der Universität Göttingen

Die Etablierung und Evaluation einer zahnärztlich-chirurgischen OSCE-Prüfung mit sechs Stationen in der ZMK-Klinik Göttingen – die Bewertung ärztlicher und studentischer Rater im Vergleich

INAUGURAL-DISSERTATION
zur Erlangung des Doktorgrades
für Zahnheilkunde

der Medizinischen Fakultät der
Georg-August-Universität zu Göttingen

vorgelegt von
Sophie-Kristin Elisabeth Schwarzer
aus
Kiel

Göttingen 2017

Dekan:	Prof. Dr. rer. nat. H. K. Kroemer
Referent/in	PD Dr. S. Sennhenn-Kirchner
Ko-Referent/in:	PD Dr. A. Simmenroth
Drittreferent/in:	……..…….

Datum der mündlichen Prüfung: Montag, 11. Dezember 2017

Hiermit erkläre ich, die Dissertation mit dem Titel "Die Etablierung und Evaluation einer zahnärztlich-chirurgischen OSCE-Prüfung mit sechs Stationen in der ZMK-Klinik Göttingen – die Bewertung ärztlicher und studentischer Rater im Vergleich" eigenständig angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Göttingen, den 23.05.2017 Sophie-Kristin Schwarzer

Inhaltsverzeichnis

Tabellen- und Abbildungsverzeichnis

Abkürzungsverzeichnis

A	-	Zahnärztlicher/ ärztlicher Rater
AEZ	-	„Aufklärung, Einwilligung Zahnextraktion“ (Station 1)
BLS	-	„Basic Life Support“ (Station 6)
Chl	-	Checklisten-Mittelwerte
d	-	Cohen-Koeffizient, Maß für die Effektstärke
EN ISO	-	Europäische Norm der International Organization for Standardization
Ex	-	Kurtosis
GB	-	Gigabyte, Maßeinheit für Datenmengen
Größe L	-	*large*, groß
Größe M	-	*medium*, mittel(groß)
Größe S	-	*small*, klein
ICC	-	Intra-Class-Correlation/ Intra-Klassen-Korrelation
ICCjust	-	justierte Intra-Class-Correlation
ICCunjust	-	unjustierte Intra-Class-Correlation
IMS	-	ItemManagementSystem
IT	-	Institut für Medizinische Informatik
LA	-	Lokalanästhesie
LAE	-	„Lokalanästhesie“ (Station 3)
M	-	Mittelwert/ Schwierigkeit
(Zahn) M	-	Molaren
MAV	-	Mund-Antrum-Verbindung
MKG	-	Mund-, Kiefer- und Gesichtschirurgie
N	-	Stichprobengröße
Naht	-	„Naht“ (Station 5)
OK	-	Oberkiefer
OP	-	Operation
OSCE	-	Objective Structured Clinical Examination
p	-	Signifikanzwert/ p-Wert
p(r)	-	p-Wert/Signifikanz für den Korrelationskoeffizienten r
PM	-	Prämolaren
POA	-	„postoperative Aufklärung“ (Station 2)
QR-Code	-	*Quick Response Code*, schnelle Antwort
r	-	Bravais- Pearson-Korrelationskoeffizient
S	-	Studentischer Rater

Sch	-	Schiefe
SD	-	Standardabweichung
SD-Karte	-	*Secure Digital Memory Card*, sichere, digitale Speicherkarte
SINUZ	-	Studentisches Innovations- und Trainingszentrum Zahnmedizin
SP	-	Schauspielpatient/ standardisierter Patient
STÄPS	-	Studentisches Trainingszentrum Ärztlicher Praxis und Simulation
UCAN	-	Umbrella Consortium for Assessment Networks
UK	-	Unterkiefer
VOSCE	-	Video-recorded Objective Structured Clinical Examination
ZARI	-	Zentrum für Anästhesiologie, Rettungs- und Intensivmedizin
ZE	-	„Zahnextraktion“ (Station 4)
ZMK	-	Zahn-, Mund- und Kieferheilkunde
α	-	Cronbachs Alpha, Reliabilitätskoeffizient

1. Einleitung

1.1 Die OSCE (Objective Structured Clinical Examination) in der Literatur

Praktische Fertigkeiten müssen auch praktisch geprüft werden, „dieses Problem wurde bereits in den 1970er-Jahren in Schottland wahrgenommen und gelöst […]“ (Simmenroth-Nayda et al. 2014, S. 235) und die erste objektivierte strukturierte klinische Prüfung [Objective Structured Clinical Examination (OSCE)] entwickelt (Davis 2003). Das Prüfungsformat OSCE beschreibt heute ein international anerkanntes und angewandtes Testverfahren, welches das Erreichen kommunikativer und praktischer Lernziele strukturiert und mit hohem Praxisbezug evaluiert (Harden und Gleeson 1979; Harden et al. 1975). Das Prinzip einer OSCE ist vergleichbar mit dem im Sportbereich angewandten Zirkeltraining. Die Studierenden absolvieren an unterschiedlichen Stationen verschiedene Aufgabenstellungen, die kurz und präzise formuliert und zu Beginn jeder Station einsehbar sind. Nach Ablauf der zur Verfügung stehenden Zeit wird an die nächste Station gewechselt und die Studierenden erhalten die Instruktion für die dortige Station. Da die Studierenden gleichzeitig die Station wechselten, konnte der Prüfungsparcours von mehreren Studierenden in Kleingruppen durchlaufen werden. Die Prüfungsthemen basieren auf den curricularen Vorgaben und werden im folgenden Abschnitt erläutert.

Über die Art der medizinischen und zahnmedizinischen Prüfungen ist bekannt, dass die am häufigsten verwendeten mündlichen und schriftlichen Prüfungsformate weniger klinische Kompetenzen und mehr Faktenwissen beurteilen. Das in der Literatur häufig zitierte und in Abbildung 1 dargestellte Modell nach Miller stellt eine Alternative dazu dar und wird in OSCE-Prüfungen umgesetzt (Nikendei und Jünger 2006; Chenot und Ehrhardt 2003). Bereits im Jahre 1990 entwickelte George E. Miller ein Schema zur Beurteilung von klinischen Fertigkeiten und Kompetenzen. Es ist demnach bekannt, dass die Beurteilung der Leistung eines Prüflings auf dem Wissen dieses Prüflings basiert („knows“), das Faktenwissen alleine jedoch nicht ausreichend zur Beurteilung klinischer Kompetenzen ist: "there is nothing more useless than a merely well informed man" (Miller 1990, S. 63). Vielmehr muss derjenige auch wissen, wie dieses Wissen anzuwenden ist („knows how“), es demonstrieren („shows how“) und dann auch praktisch anwenden können („does“). Das Schema von Miller beschreibt somit die Anforderungen, die an eine hochwertige Prüfung klinisch-praktischer Fertigkeiten gestellt werden müssen. Weder mündliche noch schriftliche Prüfungen können diese vier Säulen allein angemessen prüfen: "no single assessment method can provide all the data required for judgement of anything so complex as the delivery of professional services by a successfull physician" (Miller 1990, S. 63).

Aus der Literatur ist bekannt, dass die Anzahl an Prüfungsstationen den wichtigsten Einflussfaktor für die Qualität einer OSCE-Prüfung darstellt: Eine hochwertige OSCE-Prüfung sollte demnach 10-14 Stationen beinhalten (Nikendei und Jünger 2006).

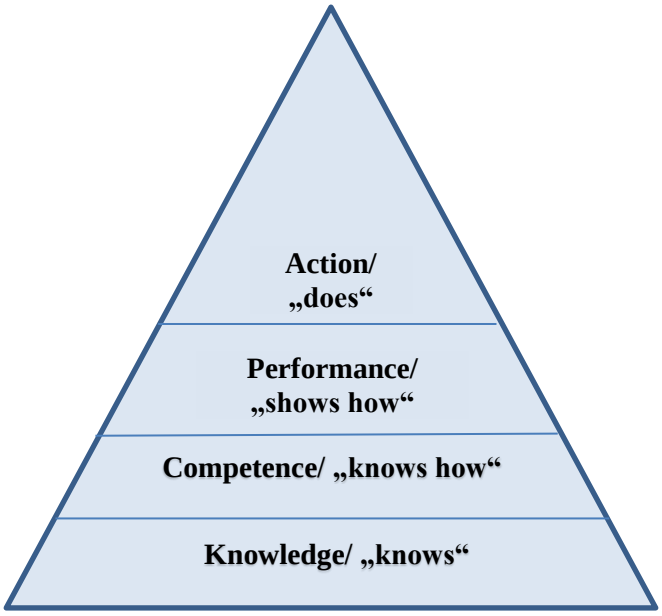

Abbildung 1: „Framework for clinical assessment“ (Miller 1990)

Schema zur Beurteilung von Fertigkeiten nach Miller: *Die vier Säulen nach Miller beschreiben die zur Prüfung klinischer Kompetenzen relevanten Aspekte: ganz unten das Faktenwissen/der Prüfling "weiß" theoretisch, was er zu tun hat (engl. „Knowledge"/ „knows"), eine Ebene darüber die Kompetenz/der Prüfling „weiß, wie" er sein Wissen ausübt (engl. „Competence"/ „knows how"), dann die Darbietung/ der Prüfling „zeigt, wie" er sein Wissen praktisch durchführt (engl. Performance/ „shows how") und schlussendlich die Durchführung/ der Prüfling „tut/demonstriert" tatsächlich sein Wissen in praktischer Form (engl. Action/ „does")* (Miller 1990, S.63).

Eine OSCE kann in zwei verschiedene Prüfungsformen unterteilt werden: Eine formative OSCE-Prüfung dient dazu, den Studierenden Defizite in ihrem aktuellen Leistungsstand aufzuzeigen und ihnen diese unter Rückmeldung in Form eines konstruktiven Feedbacks vor Augen zu führen. Sie ist primär eine Lehr-OSCE, anstelle einer Prüfungs-OSCE (Chenot und Ehrhardt 2003; Chisnall et al. 2015). Eine summative OSCE-Prüfung verfolgt hingegen eher das Ziel eine Entscheidung über die Eignung der Studierenden und über das Vorhandensein notwendiger klinischer Kompetenzen zum Bestehen eines durch das Curriculum definierten Abschnitts der zahnärztlich-chirurgischen Lehre zu treffen (Nikendei und Jünger 2006; Napankangas et al. 2014).

Durch die Etablierung der OSCE als Prüfungssystem zur Beurteilung klinischer Kompetenzen auf dem Gebiet der zahnärztlichen Chirurgie an der Universitätsmedizin Göttingen (UMG) sollen die Teilnehmer und Teilnehmerinnen ihre praktischen Fertigkeiten und Fähigkeiten an mehreren Prüfungsstationen unter Beweis stellen. Die Aufgabenstellungen beinhalten kommunikative Situationen mit ausgewählten und individuell instruierten, standardisierten Patienten und Patientinnen und praktische zahnärztlich-chirurgische Leistungen.

1.2 Überblick über das Thema

Nach erfolgreicher Pilotierung von fünf neuen OSCE-Stationen in der Zahn-, Mund-, und Kieferklinik (ZMK) der Universitätsmedizin Göttingen im Sommersemester 2014 wurde die OSCE-Prüfung am 17.10.2014 zum ersten Mal als formatives Prüfungsformat eingesetzt und im Studentischen Innovations- und Trainingszentrum Zahnmedizin (SINUZ) der ZMK-Klinik Göttingen als sechs-Stationen-OSCE durchgeführt. Zu den fünf pilotierten Stationen kam eine weitere sechste Station, die im Zentrum für Anästhesiologie, Rettungs- und Intensivmedizin (ZARI) der UMG bereits erfolgreich im Einsatz ist. Seitdem ist die Prüfung ein integraler Bestandteil des Operationskurses II in der zahnärztlich-chirurgischen Lehre der UMG. Zwei der sechs Stationen wurden mithilfe von Videoaufzeichnungen als VOSCE (Video-recorded Objective Structured Clinical Examination) ohne anwesenden Rater konzipiert und von einem studentischen und einem ärztlichen bzw. zahnärztlichen Rater sekundär ausgewertet (Kiehl et al. 2014; Hecker et al. 2012). Im Anschluss schriftlich dargelegte Selbsteinschätzungen der Studierenden wurden hierbei in Relation zu den Urteilen der Rater gesetzt (Giesler et al. 2011).

Alle Stationen wurden sowohl von einem zahnärztlichen/ärztlichen als auch von einem studentischen Rater mithilfe der Kombination von standardisierten Checklisten und der Vergabe einer Globalnote bewertet. Die Inhalte der Checklisten wurden auf Basis der curricularen Vorgaben der UMG erstellt: Die zahnärztlich-chirurgische Lehre gliedert sich an der UMG in zwei Operationskurse mit jeweils zwei Teilen, an denen die Studierenden während der klinischen Kurse teilnehmen. Beginnend im 6. Fachsemester erlernen die Studierenden im Rahmen des Operationskurses I (Teil 1) die Durchführung intraoraler Lokalanästhesien. Sowohl die Infiltrationstechnik als auch die Leitungsanästhesie werden hier gelehrt und in gegenseitigen Übungen praktisch angewandt. Außerdem wird die Nahtversorgung einer Extraktionsalveole demonstriert und am Schweinekiefer praktisch geübt. Beide Übungen werden im Zusammenhang mit den hygienischen Grundregeln dieser Behandlungsschritte gelehrt. Teil 2 des Operationskurses I erfolgt im 7. Fachsemester. Dabei wird das Legen eines intravenösen Zugangs gelehrt und ebenfalls in gegenseitigen Übungen sowie am Phantommodell angewandt. Außerdem werden parallel die Grundlagen der Zahn-Mund-Kiefer-Heilkunde

vermittelt. Im 8. Fachsemester wird die Theorie zur Zahn-Mund-Kiefer-Heilkunde vertieft und um die Grundlagen der Mund-, Kiefer- und Gesichtschirurgie erweitert. Im Rahmen des Operationskurses II (Teil 1) erleben die Studierenden den Alltag der Poliklinik und bekommen die Möglichkeit, nach vorheriger Anleitung unter Aufsicht Zähne zu extrahieren und Aufnahme- und Anamnesegespräche mit den Patienten zu führen. Der Kurs soll zukünftig mit der Teilnahme an der OSCE-Prüfung enden. Während der Fachsemester 6-9 besuchen die Studierenden die Veranstaltung „Auscultando/ Practicando“, in der sie aktuelle Patientenfälle aus der Klinik und Poliklinik für Mund-, Kiefer- und Gesichtschirurgie präsentiert bekommen und je nach Semester unterschiedliche Testatpflichtige Aufgaben absolvieren. Während die Studierenden des 6. Semesters nur zuhören, so erhebt aus den Fachsemestern 7, 8 und 9 jeweils ein Student/ eine Studentin den Röntgenbefund, die Anamnese oder führt die klinische Untersuchung durch.

Am 18. Juli 2014 wurde die OSCE in der zahnärztlichen Chirurgie mit fünf Stationen und einer Gruppe freiwilliger Studierender an der Universitätsmedizin Göttingen pilotiert (Pilot-OSCE). Es wurde außerdem der Vergleich zwischen Selbsturteil der Studierenden und Fremdurteil durch die ärztlichen und zahnärztlichen Rater erhoben. Auch die Einflüsse einzelner ärztlicher, zahnärztlicher und studentischer Rater auf die Bewertung der Prüflinge wurden im Hinblick auf eine faire Durchführung der Prüfung untersucht.

Diese Arbeit fokussiert neben Organisation und Durchführung dieser OSCE auf die statistische Auswertung und Vergleichsanalyse anhand von Video-Aufzeichnungen und Bewertungen durch ärztliche/zahnärztliche und studentische Rater sowie auf die Studierendenevaluationen. Ziel ist es dabei, klinisch relevante Situationen zu simulieren und reproduzierbar zahnärztlich-chirurgische Fähigkeiten und Fertigkeiten - integrale Bestandteile des Operationskurses II der Klinik und Poliklinik für Mund-, Kiefer- und Gesichtschirurgie - zu prüfen. Außerdem sollen die Rater-Übereinstimmungen der beiden Fremdbeurteiler evaluiert werden, um daraus Verbesserungen für die Durchführung zukünftiger OSCE-Prüfungen ableiten zu können.

1.3 Die Testgütekriterien

Die Inhalte einer OSCE-Prüfung werden anhand ihrer Testgüte bewertet, die sich aus drei Kriterien zusammensetzt: der Validität, der Reliabilität und der Objektivität. Um Aussagen über die Qualität der Testgüte dieser OSCE treffen zu können, muss geklärt werden, auf welche Art und Weise die Grundlagen der Testgütekriterien in dieser OSCE-Prüfung umgesetzt werden konnten.

1.3.1 Die Validität

Die Validität (Gültigkeit) gibt an, wie gut ein Test genau das misst, was er vorgibt zu messen. Die Gültigkeit eines Tests ist sein wichtigstes Gütekriterium (Schelten 1980; Chenot und Ehrhardt 2003).

Inhalts-, Kriteriums- und Konstruktvalidität

Die Validität setzt sich wiederum aus drei verschiedenen Aspekten zusammen. Die Inhaltsvalidität, oder Augenscheinvalidität gibt Aufschluss darüber, ob der Test das misst, was die Prüflinge gelehrt wurde. Der Test lässt Aussagen darüber zu, ob die Lernziele einer Lehreinheit durch den Test verkörpert werden und die Inhalte der curricularen Vorgaben geprüft werden. Diese Validität wird auch Augenscheinvalidität genannt und ist nicht immer numerisch bestimmbar, sondern häufig durch subjektive Wahrnehmung geprägt. Da sie nicht objektiv bestimmbar ist, sollte sie somit auch nicht als alleiniges Kriterium für die Validität herangezogen werden (Schnell et al. 1995; Schelten 1980; Chenot und Ehrhardt 2003). Die Kriteriumsvalidität ist der zweite Aspekt der Validität und wird empirisch mithilfe eines Validitäts-Koeffizienten ermittelt, der über ein externes Verfahren bestimmt und mit dem zu bewertenden Kriterium verglichen wird. Dazu wird ein Kriterium, das bereits als „valide“ definiert wurde, herangezogen. Die Korrelation dieser beiden Kriterien gibt Aufschluss über den Grad der Validität. Das Kriterium sollte dazu selbstverständlich in sinnvollem Zusammenhang stehen, also ähnliche Grunddaten ermitteln. Auf Basis der Korrelation lassen sich dann Rückschlüsse auf die Validität ziehen (Schelten 1980).

Die Konstruktvalidität gibt Aufschluss darüber, "inwieweit ein Test das abstrakte psychologische Gedankengerüst widerspiegelt, das seiner Konstruktion zugrunde liegt [...] Angst, technisches Verständnis, Intelligenz, Leistungsmotivation oder kritisches Denken sind z.B. psychologische Konstrukte" (Schelten 1980, S. 100). Die Inhaltsvalidität ist, wie bereits erwähnt, häufig subjektiv gefärbt. Die Kriteriumsvalidität ist hingegen nur selten anwendbar, da es nicht immer geeignete Vergleichsgruppen gibt, die bekannt sind und verwendet werden können. Wenn es Vergleichsgruppen gibt, so zeigen diese teilweise nur schwache Zusammenhänge an. Daher ist die Konstruktvalidität von großer Bedeutung. Sie ist aussagekräftiger als die Inhaltsvalidität und öfter anwendbar als die Kriteriumsvalidität (Hodges et al. 1998). "‚Konstruktvalidität‘ liegt dann vor, wenn aus einem Konstrukt empirisch überprüfbare Aussagen über Zusammenhänge dieses Konstruktes mit anderen Konstrukten theoretisch hergeleitet werden können und sich diese Zusammenhänge empirisch nachweisen lassen“ (Schnell et al. 1995, S. 147). Über die erhobenen Interkorrelationen zwischen den Stationen kann das Konstrukt einer OSCE-Station mit dem Konstrukt einer anderen OSCE-Station verglichen und die Zusammenhänge empirisch ermittelt werden. Jede der einzelnen Stationen stellt ein Konstrukt dar, dem die Lehrinhalte der zahnärztlichen Chirurgie aus den Operationskursen I und II zugrunde liegen, die geprüft werden sollen. Ein Konstrukt ist in diesem Fall die Aufgabenstellung

einer Station und umfasst all ihre Bewertungskriterien. Aus der Literatur ist bekannt, dass die OSCE eine hohe Augenscheinvalidität zeigt (Chenot und Ehrhardt 2003).

1.3.2 Die Reliabilität

Der Begriff der Reliabilität setzt sich aus mehreren Aspekten zusammen und gibt die Genauigkeit an, mit der etwas gemessen wird. "Bei der Reliabilität geht es um die Genauigkeit, die Zuverlässigkeit, die Treffsicherheit eines Messinstruments, unabhängig davon, ob der Test das misst, was er messen soll" (Schelten 1980, S. 102). Die Größe der Reliabilität ist somit unabhängig von der Validität, auch wenn sie häufig mit dieser gleichgesetzt wird: „In reliability, we search for the consistency or stability of the measure over time and over situations“ (Friedman und Mennin 1991, S. 390). Es ist dabei möglich, dass ein Testergebnis reliabel ist, jedoch nicht valide. Andersherum ist ein Test oder eine Prüfung, der bzw. die nicht reliabel ist auch nicht valide (Möltner et al. 2006).

Um die Reliabilität eines Tests verbessern zu können, müssen sowohl Qualitäts-, als auch Quantitätsverbesserungen der Aufgabenstellungen vorgenommen werden (Möltner et al. 2006). Zum einen kann auf die Testlänge eingewirkt werden. Je länger der Test ist, je mehr Stationen also die OSCE hat, desto geringer wird der Einfluss, den der Zufall auf die Messung hat und desto größer wird die Reliabilität der Tests und der Prüfung. Zum anderen kann durch Erhöhung der Homogenität der Teilaspekte einer Aufgabenstellung eine höhere Reliabilität erreicht werden. Für Tests mit Aufgabenstellungen aus verschiedenen Bereichen ist eine hohe Reliabilität schwieriger zu erreichen als für Prüfungen mit Aufgabenstellungen aus dem gleichen Fachbereich. Je inhaltshomogener die Testaufgaben also sind, desto größer ist auch die Reliabilität des Ergebnisses. Die dritte Möglichkeit zur Beeinflussung der Reliabilität ist die Trennschärfe bzw. Schwierigkeit eines Tests. Die genaue statistische Bedeutung wird im nächsten Abschnitt ausführlich erklärt. Die Reliabilität des Ergebnisses wird größer, je trennschärfer die zugrunde liegende Aufgabenstellung wird (Schelten 1980). Es ist daher wichtig, dass anhand der Trennschärfen Ableitungen zur Reliabilitätsverbesserung getroffen werden, also die einzelnen Items der Checklisten jeder Station auf ihre Trennschärfe überprüft werden, um zu erkennen, inwiefern eine Erhöhung der Trennschärfe für dieses Item möglich wäre. "Der Trennschärfebestimmung liegt folgender Gedanke zugrunde: Jede Aufgabe eines Testes kann als ein Test in sich aufgefasst werden. Die einzelne Testaufgabe muss das Gleiche messen, wie alle übrigen Testaufgaben zusammen" (Schelten 1980, S. 129). Die Trennschärfe zeigt also an, wie zuverlässig ein Item den Inhalt misst, den auch der Rest des Tests misst und wie gut das Item die Station repräsentiert. Werte unter 0,3 gelten als gering trennscharf und sind verbesserungsbedürftig. Sie besagen, dass „schwache“ Kandidaten dieses Item eher gut lösen und umgekehrt (Kiehl et al. 2014). Negative Werte für die Trennschärfe haben somit auch Auswirkungen auf die Reliabilität und geben an, dass

die Reliabilität der Station niedrig ist. Ein Item mit einer negativen Trennschärfe prüft also nicht das, was es vorgibt zu prüfen. Werte über 0,3 gelten als akzeptabel. Eine Trennschärfe von 0 gibt an, dass zwischen „guten“ und „schlechten“ Kandidaten nicht unterschieden wird.

Der Cronbachs-Alpha-Koeffizient gibt Aufschluss über die interne Konsistenz der Checklisten einer Station, kann aus der Korrelation der Items untereinander bestimmt werden und ist auch anwendbar, wenn dichotome Items vorliegen (Schnell et al. 1995). Präziser formuliert ist der Alpha-Koeffizient allerdings eine untere Grenze für die Reliabilität und beschreibt ein Maß für die interne Konsistenz (Möltner et al. 2006). "Ein Meßinstrument aus mehreren Indikatoren (‚Items‘) kann dann als eine Ansammlung äquivalenter Tests interpretiert werden, wenn alle Indikatoren des Instruments dieselbe Dimension messen. Diese Eigenschaft wird als ‚interne Konsistenz‘ bezeichnet." (Schnell et al. 1995, S. 142). Cronbachs Alpha gibt Aufschluss darüber, inwieweit alle Items der Checkliste einer Station dasselbe Konstrukt messen. Ein Koeffizient von 0,8 oder größer gilt dabei als akzeptabel. Das "α if deleted" lässt Aussagen über die Notwendigkeit zur Eliminierung einzelner Items aus der Checkliste zu. "Als problematisch sind dann all die Aufgaben anzusehen, bei denen das " ‚α if deleted‘ höher ist als das α der Gesamtprüfung, diese Aufgaben wirken reliabilitätsmindernd" (Möltner et al. 2006, S. 8), und eine Revision der Items sollte in Betracht gezogen werden.

1.3.3 Die Objektivität

Die Prüfung soll eine hohe Objektivität zeigen, also eine hohe Fairness und Gerechtigkeit. Ausschlaggebend ist hierbei zum einen die Interraterreliabilität (Interrater-Übereinstimmung). Die Kombination aus Bewertung durch standardisierte Checklisten und Globalnote kann die Reliabilität erhöhen (Chenot und Ehrhardt 2003; Landes et al. 2014). Die Objektivität eines Tests wird auch als Testobjektivität bezeichnet und setzte sich aus mehreren Aspekten zusammen. "Ein Test muß unabhängig von der Person sein, die ihn als Testleiter durchführt, die ihn auswertet und die die Leistungen interpretiert. Man spricht demzufolge von der Durchführungs-, Auswertungs-, und Interpretationsobjektivität." (Schelten 1980, S. 124).

Durchführungsobjektivität

Die Durchführungsobjektivität setzt voraus, dass die äußeren Bedingungen unter denen eine Prüfung stattfindet gleich sind und die Anweisungen zu deren Durchführung detailliert festgehalten werden, um differenzierte Auslegungen der Durchführungseigenschaften zu vermeiden. Dazu gehört eine Festlegung des zeitlichen Rahmens des Testumfangs, der zu benutzenden Hilfsmittel und Hilfestellungen durch andere Personen, der Verbote und Pflichten der Teilnehmer und der Aufgabenstellungen. Werden diese Anweisungen strikt durchgesetzt, so entstehen standardisierte Teilnahmebedingungen und die Durchführungsobjektivität gilt als erfüllt (Schelten 1980).

Auswertungsobjektivität

Kann ein Testergebnis von jeder beliebigen Person erhoben werden, so sind die Voraussetzungen für die Auswertungsobjektivität geschaffen. Das Auswertungsergebnis ist dabei unabhängig von der bewertenden Person. Dies wird dadurch ermöglicht, dass die Bewertungskriterien keinen Interpretationsspielraum zulassen und genau festlegen, was diese beliebige Person als „richtig“ und was als „falsch“ zu bewerten hat. Die Auswertungsobjektivität ist am besten zu verwirklichen, wenn die Auswertung auf einer Richtig-Falsch-Basis, Single-Choice-Item beruht (Chenot et al. 2007). So wird der Bewertungsspielraum minimiert. Bereits eine Bewertung über eine Differenzierung in „gut“, „mittel“, „schlecht“, eine sogenannte „likert scale“ ergibt eine im Vergleich dazu erschwerte Auswertungsobjektivität, da der bewertenden Person mehr Spielraum für Interpretationen gewährt wird (Chenot et al. 2007). Dennoch sind beide Arten der Auswertung sehr gut geeignet, denn bei "Richtig-Falsch Aufgaben oder Mehrfachwahlaufgaben mit einer eindeutigen Lösung ist die Auswertungsobjektivität praktisch vollkommen verwirklicht" (Schelten 1980, S. 126). Dies kann mithilfe von Checklisten verwirklicht werden, die die Auswertungskriterien klar definieren und diese beiden Aufgabentypen beinhalten.

Interpretationsobjektivität

Der dritte Aspekt der Testobjektivität ist die Interpretationsobjektivität. Diese ist gegeben, "wenn zwei Beurteiler unabhängig voneinander aus dem gleichen Auswertungsergebnis den gleichen Schluß ziehen" (Schelten 1980, S. 126). Sie setzt voraus, dass das gleiche Ergebnis eines Tests von zwei Beurteilern unabhängig voneinander auf die gleiche Art und Weise interpretiert werden kann. Für nicht-standardisierte Tests ist dies aussagekräftiger als für standardisierte Tests, da für standardisierte Tests der Maßstab auf dem der Vergleich beruht gleich ist, und somit eine Auswertung der Ergebnisse zwangsläufig zur gleichen Interpretation führt (Schelten 1980).

1.4 Erläuterungen zu statistischen Parametern

1.4.1 Die Itemanalyse

Die im Folgenden erläuterten statistischen Größen sind in Tabelle 1 zur verbesserten Übersicht noch einmal zusammengefasst dargestellt.

Die Berechnung der Mittelwerte setzt sich aus den Ergebnissen aller Studierenden zusammen, dividiert durch die Anzahl der Stichprobe N. Die Schwierigkeit gibt an, wie viele Kandidaten die Aufgabe richtig gelöst haben. Die Punktzahlen können auch in Prozentzahlen interpretiert werden: 100% = 1 = alle; 50% = 0,5 = die Hälfte. Dementsprechend bedeutet eine Schwierigkeit von 0, dass niemand das Item richtig gelöst hat, wohingegen eine Schwierigkeit von 1 angibt, dass alle Kandidaten die Aufgabe richtig gelöst haben. Eine Schwierigkeit von 1 bedeutet folglich, dass die Aufgabe leicht war, eine Schwierigkeit von 0, dass sie schwer war. Nimmt die Schwierigkeit den Wert von 0,5 an, so kann sie als mittelschwer bezeichnet werden (Schelten 1980). Dies ist auf die Items der Checklisten übertragbar. Mithilfe der Itemanalyse ist es möglich, für jedes Item eine Aussage darüber zu treffen, wie gut es in der Lage ist zwischen leistungsstarken und leistungsschwachen Kandidaten zu unterscheiden.

"Aufgaben, die von nahezu allen oder von nahezu niemandem gelöst werden, tragen zur Differenzierung zwischen den leistungsstarken und den leistungsschwachen Personen nichts bei. In einem möglichst hoch differenzierenden Test müssen die extrem leichten und extrem schweren Aufgaben eliminiert werden" (Schelten 1980, S. 129).

"Als Richtwert für die Aufgabenschwierigkeit wird in der Literatur der Bereich von etwa 0,4 bis 0,8 empfohlen" (Möltner et al. 2006, S. 3). Aufgaben mit einer Schwierigkeit oberhalb dieses Richtwerts werden als zu leicht definiert, Aufgaben mit Schwierigkeiten unterhalb dieses Bereichs als zu schwer: "aus Sicht der klassischen Testtheorie, in der vor allem der Aspekt einer guten Differenzierung betont wird, sind solche Aufgaben in Hinblick auf die Prüfungsökonomie überflüssig" (Möltner et al. 2006, S. 3). Die Trennschärfe wird folglich für sehr große und sehr kleine Schwierigkeitswerte niedrig und ist somit ebenfalls nicht geeignet, um Aussagen über den Grad der Leistung der Prüflinge zu treffen.

Die Erläuterungen zur Steigerung der Testgüte werden noch zeigen, dass es auch sinnvoll sein kann, basale Kenntnisse abzufragen, unabhängig davon, ob ihre zu erwartende Schwierigkeit im empfohlenen Bereich liegt.

Die Berechnung der Standardabweichung gibt Informationen über die Streuung der Daten, also über die durchschnittliche Entfernung der jeweiligen Messpunkte zum Mittelwert. Die Trennschärfe zeigt an, wie gut ein einzelnes Item die jeweilige OSCE Station repräsentiert. Negative Trennschärfen sind ungünstig, da sie besagen, dass Prüflinge, die dieses Item schlecht gelöst haben

insgesamt gut waren und umgekehrt. Dies drückt Zweifel an der Verständlichkeit der Aufgabenstellung der Station aus oder kann bedeuten, dass die Auswertung nicht richtig gepolt wurde. Genauere Erläuterungen dazu folgen. Wünschenswert sind Trennschärfen von über 0,3. Die Trennschärfe erlaubt unter anderem eine Aussage über die Reliabilität (Messgenauigkeit) der Prüfung. "Die Zahl der Stationen spielt auch für die Reliabilität eine Rolle. Die Variabilität der Leistung eines Prüflings an verschiedenen Stationen (interstation reliability) wird als Reliabilitätskoeffizient (Alphakoeffizient von Cronbach) ausgedrückt" (Chenot und Ehrhardt 2003, S.4), auch Cronbachs Alpha genannt. Die Messgenauigkeit, ausgedrückt durch Cronbachs Alpha, kann Werte zwischen 1,0 und 0,0 annehmen, wobei 1,0 eine absolute Messgenauigkeit beschreibt, Werte um 0,5 eine mittelmäßige Messgenauigkeit und Werte um 0,00 absolut keine Messgenauigkeit darstellen. Bereits ein Alpha-Koeffizient von 0,3 stellt ein zu geringes Ergebnis dar. „Ab einem Wert von 0.8 geht man von einer guten Reliabilität aus“ (Chenot und Ehrhardt 2003, S. 4), die für *high stakes* Examina gefordert wird (Nikendei und Jünger 2006). Die Reliabilität gibt sowohl Auskunft über „die Zuverlässigkeit oder Reproduzierbarkeit von Prüfungsergebnissen“ (Möltner et al. 2006, S. 7), als auch über die "Fähigkeit einer Prüfung, kompetente von nicht-kompetenten Studierenden zu unterscheiden (Sensitivität, Spezifität)" (Chenot und Ehrhardt 2003, S. 3).

Im Rahmen der Zusammenhangsanalyse wurden die Werte für die Korrelation zwischen Skalenmittelwert (Checklisten-Mittelwert) und Note (Globalnote) ermittelt. Die Korrelation zeigt die Enge des Zusammenhangs zweier Merkmale an und wird numerisch definiert durch den Korrelationskoeffizienten „r“, der Werte zwischen +1 und -1 annehmen kann. Der Wert 1 steht dabei für einen hohen Zusammenhang, während 0 keinen Zusammenhang beschreibt.

"Erreicht ein Korrelationskoeffizient Werte von +1 bzw. -1, geht der stochastische Zusammenhang in einen funktionalen, deterministischen Zusammenhang über, wobei eine Korrelation von +1 einen linearen gleichsinnigen Zusammenhang und eine Korrelation von -1 einen linearen, gegenläufigen Zusammenhang anzeigt" (Bortz 1985, S. 213–214).

Zusätzlich zu Cronbachs Alpha für die Gesamtheit der Station und das jeweilige Item wurde noch ein weiterer Alpha-Koeffizient für jedes Item berechnet. Dieser Wert wird auch als "α if deleted" (Möltner et al. 2006, S. 8) bezeichnet. Er gibt genau den Alpha-Koeffizienten an, der sich für die gesamte Station ergäbe, wenn das Item eliminiert werden würde. Dieser Wert wird bestimmt, um Aussagen über die Zuverlässigkeit der Messung unter Ausschluss der jeweiligen Items treffen zu können. Zuletzt wird die Korrelation zwischen der Note und dem Skalenmittelwert angegeben.

Tabelle 1: Übersicht der statistischen Parameter

Name	**Aussage**	**Dimension**
Schwierigkeit/ Mittelwert (M)	Wie viele Studierende haben die Aufgabe richtig gelöst?	1,0 = 100% = alle = leicht; 0,5 = 50 % = die Hälfte = mittelschwer; 0,0 = 0 % = keiner = schwer
Standardabweichung (SD)	Streuung der Daten, durchschnittliche Entfernung der Messpunkte zum Mittelwert	Quadratwurzel aus der Varianz $0 \geq SD \geq +\infty$
Trennschärfe	Wie gut repräsentiert ein Item die Station?	> 0,3 = gut < 0,3 = nicht gut < 0,0 = Fehler in Aufgabenstellung oder Polung der Auswertung
Cronbachs Alpha, Reliabilitätskoeffizient, Alphakoeffizient von Cronbach	Reliabilität des Items	≥ 0,8 = gute Reliabilität ≤ 0,3 = geringe Reliabilität
„Alpha if deleted“/ „α if deleted“	Wert, der sich als Gesamt-Alpha über die Station ergibt, wenn dieses Item eliminiert werden würde	siehe Cronbachs Alpha
Messgenauigkeit	Wie genau ist die Messung?	1,0 = absolute Messgenauigkeit 0,5 = mittelmäßige Messgenauigkeit 0,0 = keine Messgenauigkeit
Korrelationskoeffizient r	beschreibt den Zusammenhang zwischen zwei Merkmalen	Quotient aus der Kovarianz der beiden Variablen und dem Produkt der Standardabweichung dieser Variablen +1 = hohen, gleichsinnigen Zusammenhang 0 = keinen Zusammenhang -1 = hohen, gegenläufigen Zusammenhang

Schiefe (Sch)	Differenz zwischen dem arithmetischen Mittel und dem Modalwert, dividiert durch die Standardabweichung	0 = normalverteilte Funktion > 0 = linkssteil, rechtsschief < 0 = rechtssteil, linksschief -0,5 > Sch > 0,5 = annährend normalverteilte Funktion (akzeptabler Bereich)
Kurtosis (Ex)	Interquartilbereich dividiert durch den Interdezilbereich, welcher mit dem Faktor 2 multipliziert wird; Wölbung	0 = mesokurtisch, mittelmäßig gewölbt, normalverteilte Funktion; > 0 = leptokurtisch, stark gewölbt, schmalgipflig; < 0 = platykurtisch, flach gewölbt, breitgipflig; -1 > Ex > +1 = annährend normalverteilte Funktion (akzeptabler Bereich)
Minimum (Min)	Unteres Ende der Verteilung, geringster Wert	$-\infty >$ Min $> +\infty$
Maximum (Max)	Oberes Ende der Verteilung, höchster Wert	$-\infty >$ Max $> +\infty$
Modalwert (Mo)	Der Wert, der am häufigsten besetzt ist; graphisch gesehen, der Wert, bei dem die Verteilung ihr Maximum hat	$-\infty <$ Mo $< +\infty$
Interdezilbereich	Streubreite der mittleren 80% aller Punkte	Differenz zwischen dem neunten und dem ersten Dezil (Zehntel)
Interquartilbereich	die Streubreite der mittleren 50% aller Punkte	Differenz zwischen dem dritten und ersten Quartil (Viertel)
t-Verteilung	gibt die Wahrscheinlichkeit an, mit der die ermittelten Mittelwertsdifferenzen zwischen allen rein theoretisch	

	denkbaren Differenzen auftreten; sie beschreibt somit die Zufälligkeit der Übereinstimmung der beiden abhängigen Stichproben	
Freiheitsgrade (df)	gibt die Anzahl an Werten an, die in einer Berechnungsformel frei variieren dürfen	Stichprobengröße N - 1 für abhängige Stichproben; (N1 + N2) - 2 für unabhängige Stichproben
t-Wert	Standardisierter Stichprobenerkennwert	Differenz zwischen empirischer und theoretischer Mittelwertedifferenz dividiert durch den geschätzten Standardfehler der Mittelwertedifferenz
p-Wert	Signifikanzniveau; gibt die Wahrscheinlichkeit an, mit der die Differenz als Folge eines Zufalls bzw. als Folge des Standardfehlers entstanden ist	Kann Werte zwischen 0 und 1 annehmen $p < 0,05$ = Ergebnis signifikant = Ergebnis ist nicht zufällig entstanden; $p < 0,01$ = Ergebnis sehr signifikant; $p < 0,001$ Ergebnis höchst signifikant; $0,2 > p > 0,05$ = Indifferenzbereich, beide Interpretationen möglich; $p > 0,2$ = Ergebnis ist nicht signifikant = Ergebnis ist zufällig entstanden
Standardfehler	Standardabweichung der Mittelwerte von gleichgroßen Zufallsstichproben einer Population	Wurzel aus der Varianz

Effektstärke	lässt Rückschlüsse auf die praktische Bedeutsamkeit eines statistisch signifikanten Ergebnisses zu	Darstellung mithilfe von Cohen-Koeffizient d
Cohen-Koeffizient d	beschreibt die Effektstärke	Mittelwertedifferenz dividiert durch die gemeinsame Varianz; Kann Werte zwischen -1 und 1 annehmen -1 = minimale Effektstärke; 0 = keine Effektstärke; +1 = maximale Effektstärke; $d \geq 0,2$ = schwacher Effekt; $d \geq 0,5$ = mittelstarker Effekt; $d \geq 0,8$ = starker Effekt
ICC-Koeffizient (Intra-Class-Correlation, Intra-Klassen-Korrelation)	Korrelationskoeffizient, mit dessen Hilfe Aussagen über die Reliabilität von Messwertereihen getroffen werden und eine Abschätzung des Zusammenhangs gepaarter Beobachtungen gegeben werden kann	0 = zufälliges Bewertungsurteil 1 = perfekt zuverlässiges Bewertungsurteil $0,4 \leq ICC \leq 0,75$ = gute/ mittelmäßige Reliabilität $ICC \geq 0,75$ = exzellent $ICC \leq 0,4$ = schlecht

1.4.2 Deskriptive Statistik und Verteilungseigenschaften

Für die Checklisten-Mittelwerte, die Noten und die Gesamtbewertung wurden im Rahmen der deskriptiven Statistik der Mittelwert (M), die Standardabweichung (SD), die Schiefe (Sch), die Kurtosis (Ex) sowie das Minimum (Min) und Maximum (Max) dargestellt. Der Mittelwert wurde rekrutiert aus den Mittelwerten der einzelnen Items einer Station. Verglichen werden die Verteilungseigenschaften mit den Merkmalen einer Normalverteilung. Diese wird beschrieben als eine Verteilung, die „unimodal und symmetrisch ist, und zudem annährend einen glockenförmigen Verlauf aufweist“ (Bortz 1985, S. 47). Eine Normalverteilung wird dadurch gekennzeichnet, dass zwei Drittel (68,26%) der erhobenen Messwerte in dem Bereich zwischen dem Mittelwert plus/minus dem Wert einer Standardabweichung liegen: M – 1 SD bis M + 1 SD. In dem Bereich von M – 2 SD bis M + 2

SD befinden sich im Rahmen einer Normalverteilung 95% der Messwerte (Bortz 1985; Sachs und Hedderich 2009). Es ist wichtig zu ermitteln, inwieweit sich im Rahmen dieser Studie normalverteilte Daten ergeben, da dies Voraussetzung zur statistischen Analyse der Mittelwertedifferenzen zwischen den Ratergruppen ist. Um den t-Test anwenden zu können, müssen die zu untersuchenden Merkmale normalverteilt sein (Rasch et al. 2014b; Rumsey 2010). Für Normalverteilungen lassen sich einige allgemeine Aussagen formulieren: Die Parameter zur theoretischen Verteilung der Grundgesamtheit einer Normalverteilung werden anhand der empirischen Momente beschrieben (Bärlocher 2008; Schlosser 1976; Bortz 1997).

Das erste empirische Moment entspricht dem arithmetischen Mittelwert, das zweite der empirischen Varianz. Die Schiefe wird als das dritte empirische Moment bezeichnet. Die Berechnung der Schiefe empirischer Häufigkeitsverteilungen setzt sich zusammen aus der Differenz zwischen dem arithmetischen Mittel und dem Modalwert, dividiert durch die Standardabweichung. „Der Modalwert (Mo) einer Verteilung ist derjenige Wert, der am häufigsten besetzt ist bzw. in der graphischen Darstellung einer Verteilung der Wert, bei dem die Verteilung ihr Maximum hat.“ (Bortz 1985, S. 47). Entspricht die Schiefe dem Wert von 0, so ist die Verteilung symmetrisch, also normalverteilt. Ein positiver Wert für die Schiefe bedeutet, dass die Verteilung linkssteil (oder rechtsschief) ist, ein negativer Wert, dass die Verteilung rechtssteil (oder linksschief) ist (Bortz 2005). Die Kurtosis - auch Exzess oder Wölbung genannt - ist das vierte empirische Moment und errechnet sich aus den Quartil-, bzw. Dezilwerten. Die Formel dafür entsteht aus dem Interquartilbereich, dividiert durch den Interdezilbereich, welcher mit dem Faktor 2 multipliziert wird (Bortz 1985). Der Interdezilbereich beschreibt die "Streubreite der mittleren 80% aller Fälle" (Bortz 1985, S. 54), während der Interquartilbereich die Differenz zwischen dem dritten und ersten Quartil darstellt, also die mittleren 50% der Streubreite. Die Kurtosis lässt Aussagen über die Steilheit einer Verteilung zu. Sie ist mesokurtisch (mittelmäßig gewölbt), wenn sie dem Wert 0 entspricht, oder normalverteilt. Nimmt sie negative Werte an, so wird sie als platykurtisch (flach gewölbt) oder breitgipflig beschrieben, bei positiven Werten als leptokurtisch (stark gewölbt) oder schmalgipflig (Bortz 1985; Sachs und Hedderich 2009). Annährend normalverteilt sind Daten, deren Schiefe und Kurtosis nahe Null liegen. Der akzeptable Bereich liegt dabei zwischen - 0,5 und + 0,5 für die Schiefe und zwischen -1 und +1 für die Kurtosis. Alle Werte außerhalb dieses Bereichs können nicht mehr als normalverteilt bezeichnet werden. Minimum und Maximum geben Auskunft über die beiden Enden der Verteilung.

1.4.3 Mittelwertevergleiche mit t-Tests

1.4.3.1 Abhängige Stichproben

Für die Mittelwertevergleiche wurde die t-Verteilung der Bewertung der zahnärztlichen und studentischen Rater erhoben (N = 36). Die t-Verteilung oder standardisierte Verteilung gibt die Wahrscheinlichkeit an, mit der die ermittelten Mittelwertedifferenzen zwischen allen rein theoretisch denkbaren Differenzen auftreten. Sie beschreibt somit die Zufälligkeit der Übereinstimmung der beiden abhängigen Stichproben. Stichproben werden als abhängig bezeichnet, „wenn an einer Stichprobe zwei Messungen durchgeführt werden (Meßwiederholung)“ (Bortz 1985, S. 170). "Verbundene (oder abhängige) Stichproben haben immer denselben Umfang [...] Jeder Wert der einen Stichprobe bildet mit einem Wert der anderen Stichprobe inhaltlich ein Paar" (Weiß 2013, S. 175). In diesem Fall werden die Leistungen der Studierenden unter der Bewertung der zahnärztlichen Rater mit der Bewertung durch die studentischen Rater verglichen. Die Form der t-Verteilung ist dabei abhängig von der Stichprobengröße (N) und der Anzahl der Freiheitsgrade (df). "Die Anzahl der Freiheitsgrade gibt an, wie viele Werte in einer Berechnungsformel frei variieren dürfen, damit es zu genau einem bestimmten Ergebnis kommt" (Rasch et al. 2014a, S. 54). Die Anzahl der Freiheitsgrade errechnet sich für abhängige Stichproben aus der Größe der Stichprobe N - 1. Die t-Verteilung ist im Vergleich zur Standardnormalverteilung schmalgipfliger (Kurtosis > 0) und flacher. Je größer die Stichprobe (N, und folglich df) wird, desto mehr zeigt die t-Verteilung Eigenschaften einer Standardnormalverteilung. Bei $N = \infty$ geht die t-Verteilung in eine Normalverteilung über, bei $N < \infty$ ist die t-Verteilung schmalgipfliger und flacher. Das gleiche gilt für die Anzahl der Freiheitsgrade. Je kleiner die Anzahl an Freiheitsgraden, desto flacher wird die Verteilung und desto größer wird die Wahrscheinlichkeit für einen bestimmten t-Wert. Der t-Wert, oder standardisierter Stichprobenkennwert entsteht mathematisch aus der Differenz zwischen empirischer und theoretischer Mittelwertedifferenz, dividiert durch den geschätzten Standardfehler der Mittelwertedifferenz (Rasch et al. 2014b). Der Standardfehler gibt Aufschluss über die Streuung der Stichprobenwerte vom arithmetischen Mittelwert und ist definiert als die „Standardabweichung der Mittelwerte von gleichgroßen Zufallsstichproben einer Population“ (Bortz 1985, S. 56). Mathematisch gesehen entspricht er der Wurzel aus der Varianz und nimmt mit steigender Varianz zu und mit steigender Stichprobengröße ab. Je größer der t-Wert (T) wird, desto signifikanter ist das Ergebnis. Das Signifikanzniveau wird mithilfe des p-Werts angegeben. Ein kleiner t-Wert mit zugehörigem hohen p-Wert gibt an, dass die Mittelwerteunterschiede nicht-signifikant sind, also zufällig entstanden sind und keiner Systematik zugrunde liegen. Ein hoher r-Koeffizient zeigt hingegen, dass die Verteilung positiv korreliert und der p-Wert von r, im Folgenden als p(r) bezeichnet, Aufschluss über die Signifikanz des Korrelationsergebnisses gibt.

Der p-Wert gibt in Kombination mit dem t-Wert Aufschluss über die Signifikanz der erhobenen Stichprobendifferenz. Er kann Werte zwischen 0 und 1 annehmen und als prozentuale Angabe interpretiert werden: z. B. 0,03 entspricht 3%. Der p-Wert gibt die Wahrscheinlichkeit an, mit der die Differenz als Folge eines Zufalls, oder anders ausgedrückt als Folge des Standardfehlers auftritt. Ist der p-Wert sehr niedrig, so bedeutet dies, dass das gewonnene Ergebnis nicht zufällig entstanden ist, sondern "ein systematischer Effekt" (Rasch et al. 2014a, S. 56) existiert. Diese Erkenntnis kann dazu führen, dass die eigene Hypothese bestärkt oder verworfen wird. Die Wahl eines Signifikanzniveaus ermöglicht es einen kritischen p-Wert und folglich einen kritischen t-Wert zu bestimmen und so eine empirische Einschätzung über die Bedeutung des t-Wertes geben zu können. Ist der p-Wert kleiner als 0,05, so gilt das Ergebnis als signifikant. Für diesen speziellen t-Test würde dies bedeuten, dass die Wahrscheinlichkeit dafür, dass die erhobene Mittelwertedifferenz nicht zufällig entstanden ist, unter 5% liegt. Ein p-Wert unter 0,01 bzw. 0,001 gilt als sehr bzw. höchst signifikant und schließt einen Zufall nahezu aus. "Diese Entscheidung ist allerdings nie zu 100% sicher, denn jede empirische Mittelwertsdifferenz ist prinzipiell auch unter der Nullhypothese möglich" (Rasch et al. 2014a, S. 56). Im Bereich $0{,}2 > p > 0{,}05$ liegt der Indifferenzbereich. Das Ergebnis dieser Werte ist kein eindeutiges Zeichen für das Verwerfen oder das Bestätigen der Hypothese und kann in beide Richtungen interpretiert werden. Nimmt p einen Wert an, der größer als 0,2 ist, so ist die Wahrscheinlichkeit dafür, dass das Ergebnis zufällig entstanden ist groß. Es kann nicht davon ausgegangen werden, dass den Mittelwertedifferenzen eine Systematik zugrunde liegt und das Ergebnis ist nicht signifikant.

Mithilfe der Korrelation kann ermittelt werden, wie der Zusammenhang der beiden untersuchten Merkmale zu beschreiben ist. Die Korrelation setzt sich zusammen aus dem Korrelationskoeffizienten r und dem dazugehörigen p-Wert. Für normalverteilte Daten mit linearem Zusammenhang wird die Bravais-Pearson-Korrelation, oder auch Produkt-Moment-Korrelation genannt, verwendet (Cohen 1992). Der Korrelationskoeffizient ist, mathematisch betrachtet, der Quotient aus der Kovarianz der beiden Variablen und dem Produkt der Standardabweichung dieser Variablen (Bortz 1985). Der zugehörige p-Wert gibt die Signifikanz dieses Wertes an.

Die Produkt-Moment-Korrelation ist ein *Inter*klassenkorrelationsmaß, d.h. die Korrelation besitzt die Eigenschaft, den Zusammenhang zwischen zwei Merkmalen zu analysieren "[...] die eine unterschiedliche Metrik besitzen: Es können Merkmale mit unterschiedlichen Maßeinheiten und Varianzen zueinander in Beziehung gesetzt werden." (Wirtz und Caspar 2002, S. 157–158). Auf die *Intra*klassenkorrelation wird im Ergebnisteil noch genauer eingegangen. Der Korrelationskoeffizient kann Werte zwischen -1 und +1 annehmen. Ist $r = 0$, so lässt sich kein Zusammenhang zwischen den beiden Variablen erkennen und sie werden als unkorreliert bezeichnet. Nimmt „r“ einen positiven Wert an, so spricht man von einem positiven Zusammenhang zwischen den beiden Variablen, das bedeutet größere Werte der einen Variable gehen auch mit größeren Werten der anderen Variable

einher. Ist $r < 0$, so tritt genau der entgegengesetzte Fall ein und die Variablen gelten als negativ korreliert (Bortz 1985).

1.4.3.2 Unabhängige Stichproben

Mithilfe des t-Tests für unabhängige Stichproben lässt sich ermitteln, ob zwischen zwei unterschiedlichen Stichproben ein Zusammenhang besteht, der statistisch signifikant ist. Die Umfänge der Stichprobe können bei dieser Art von Stichprobentest ungleich sein. Die Größe der Freiheitsgrade (df) errechnet sich „aus normalverteilten Grundgesamtheiten mit $n_1 + n_2 – 2$“ (Sachs und Hedderich 2009, S. 437). Für den Vergleich der zahnärztlichen Rater untereinander sowie der studentischen Rater untereinander ergibt sich aufgrund der beiden gleich großen Stichprobengruppen von $N_1 = 18$ und $N_2 = 18$ somit eine Anzahl an Freiheitsgraden von 34 (df = 34). Das bedeutet, dass zwei unterschiedliche Stichproben mit je 18 Messwerten von zwei unterschiedlichen Ratern verglichen werden, denen jedoch dieselben Bewertungskriterien zugrunde liegen. In die Berechnung des t-Werts geht mathematisch betrachtet der Standardfehler der Mittelwertedifferenz ein, welcher von der jeweiligen Varianz abhängt. Da jede Varianz die Freiheitsgradzahl N-1 beinhaltet ergeben sich für die Stichprobe die Formel N_1-1 + N_2-1 oder anders ausgedrückt $N_1 + N_2 – 2$ Werte, die frei variieren können (Rasch et al. 2014b; Weiß 2013).

Die Bestimmung der Signifikanz mithilfe des p-Werts ist abhängig von der Stichprobengröße bzw. der Streuung und wird mit steigender Stichprobengröße oder abnehmender Streuung signifikanter. Ein von der Stichprobengröße unabhängiges Maß stellt die Effektstärke dar. Sie lässt Rückschlüsse auf die praktische Bedeutsamkeit eines statistisch signifikanten Ergebnisses zu. Sie gibt also an, wie bedeutsam die Unterschiede zwischen zwei verschiedenen Gruppen sind. "Die Größe eines empirischen Effekts ist für die inhaltliche Bewertung eines signifikanten Ergebnisses wichtig, da durch eine Erhöhung des Stichprobenumfangs theoretisch jeder noch so kleine Effekt signifikant gemacht werden kann." (Rasch et al. 2014a, S. 48). Die Effektstärke kann mithilfe der standardisierten Differenz, dem Cohen-Koeffizienten d dargestellt werden. Dieser errechnet sich aus der Mittelwertedifferenz dividiert durch die gemeinsame Varianz (Sachs und Hedderich 2009). Er ist ein direktes Effektmaß für Mittelwertevergleiche gleich großer Gruppen und kann Werte zwischen -1 und +1 annehmen. Der Wert +1 bezeichnet dabei eine maximale Effektstärke, 0 keine Effektstärke und -1 eine minimale Effektstärke. Ab einer Effektstärke von d = 0,2 spricht man von einem geringen/schwachen Effekt, ab d = 0,5 von einem mittelgroßen Effekt und ab d = 0,8 von einem großen/starken Effekt (Sachs und Hedderich 2009; Raupach et al. 2011). Das Effektmaß wird im Unterschied zu dem Korrelationskoeffizienten r nicht darauf beschränkt lineare Zusammenhänge zwischen Mittelwertedifferenzen darzustellen, sondern auch nicht-lineare.

1.4.4 Intra-Klassen-Korrelation (ICC)

Die Intra-Class-Correlation (ICC), oder Intra-Klassen-Korrelation, beschreibt eine Gruppe von Korrelationskoeffizienten, mit deren Hilfe Aussagen über die Reliabilität von Messwertereihen getroffen werden und eine Abschätzung des Zusammenhangs gepaarter Beobachtungen gegeben werden kann. Sie ist ein Maß für "die Stärke des Zusammenhangs der Urteile zweier Rater, die dieselben Personen oder Objekte beurteilt haben [...] Anders als bei üblichen Korrelationsmaßen kann zwischen einer unjustierten und einer justierten ICC gewählt werden." (Wirtz und Caspar 2002, S. 189). Der Begriff der *Inter*korrelation wurde bereits verwendet. Im Unterschied dazu ist die *Intra*klassenkorrelation "ein strengeres Maß, weil hier die Bedingung erfüllt sein muss, dass für dasselbe Merkmal mehrere Messwertreihen vorliegen: Bei der Frage nach der Interraterreliabilität wird beispielsweise dasselbe Merkmal durch mehrere Rater eingeschätzt." (Wirtz und Caspar 2002, S. 158).

Werden die Testpersonen, in diesem Fall die studierenden Prüflinge, nicht alle von denselben Ratern beurteilt, so spricht man von einer unjustierten ICC. Die der ICC zugrunde liegende Varianzanalyse wird folglich als einfaktoriell bezeichnet. Zweifaktorielle Varianzanalysen entstehen hingegen, wenn alle Studierenden von denselben Ratern beurteilt werden. Das zugehörige Maß, die ICC, kann sowohl unjustiert, als auch justiert sein (Wirtz und Caspar 2002). Die beiden Videostationen der OSCE ergeben folglich eine zweifaktorielle Varianzanalyse, da die Bewertung von einem Rater durchgeführt wurde und somit derselbe Rater alle Studierenden bewertete. Die anderen vier Stationen wurden von mehr als einem Rater beurteilt. Dabei bewertet ein Rater nicht alle, sondern einen Teil der Studierenden, weshalb die Varianzanalyse in diesem Fall als einfaktoriell bezeichnet werden muss und die zugehörige ICC unjustiert ist.

Ein ICC-Wert von 0 zeigt ein zufälliges Beurteilungsergebnis an und sagt somit aus, dass kein Zusammenhang zwischen den Urteilen der Rater erkennbar ist. Ein ICC-Wert von 1 stellt hingegen eine perfekt zuverlässige Merkmalseinschätzung durch den Beurteiler dar. Je mehr sich der ICC-Wert diesem Bereich nährt, desto höher kann die Reliabilität der Rater bewertet werden und desto höher ist der Zusammenhang zwischen den beiden Urteilen. "Allgemein wird in der Literatur eine Intraklassenkorrelation von mindestens 0,7 als Indiz für eine "gute" Reliabilität angesehen“ (Wirtz und Caspar 2002, S. 160). Höhere Werte als 0,75 werden als „exzellent“ bezeichnet, Werte zwischen 0,4 und 0,75 als „gut“ oder „akzeptabel“ und Werte unter 0,4 als „schlecht“ (Kiehl et al. 2014).

1.5 Formulierung der Fragestellung und der Hypothesen

Ziel der Studie war es zum einen, die sechs Stationen auf ihre **Testgüte** zu überprüfen, um sie nach Überarbeitung in Zukunft als Teil einer summativen OSCE-Prüfung etablieren zu können. Zum anderen sollte die **Rater-Übereinstimmung** zwischen den beiden Urteilen der zahnärztlichen/ ärztlichen und der studentischen Rater untersucht werden, um gegebenenfalls zur Unterstützung der zahnärztlichen Lehre ärztliche Ressourcen schonen zu können und formative OSCE-Prüfungen und Pilotierungen weiterer Stationen mit studentischen Ratern anstelle von ärztlichen/ zahnärztlichen Ratern durchführen zu können. Die Fragestellung nach der Ersetzbarkeit von approbierten ärztlichen und zahnärztlichen Ratern durch studentische Rater als alternative Bewertungsmöglichkeit der OSCE-Stationen setzt voraus, dass beide Untersuchungsgruppen zuvor an einer ausführlichen Schulung teilnehmen (Chenot et al. 2007). Die Leistungen der Teilnehmenden wurden von den Ratern mithilfe von Checklisten und Globalnoten bewertet, aus denen sich das Gesamtergebnis zusammensetzte. Der Inhalt der sechs Stationen prüfte Grundlagen der zahnärztlich-chirurgischen Praxis, kommunikative Fähigkeiten und Durchführung des Basic-Life-Supports als lebensrettende Sofortmaßnahme im Rahmen der Ersten Hilfe.

Für den Vergleich der Bewertung durch die zahnärztlichen und die studentischen Rater können zwei Hypothesen aufgestellt werden: Die Nullhypothese, der im Rahmen dieser Arbeit nachgegangen wird, geht davon aus, dass es keine signifikanten Mittelwerteunterschiede zwischen den ärztlichen Ratern (A) und den studentischen Ratern (S) gibt. Das bedeutet, dass beide Gruppen die Studierenden im Hinblick auf die Checklisten-Mittelwerte und die Globalnote gleich bewerten und sich für das Gesamtergebnis gleiche Mittelwerte ergeben. Die Alternativhypothese H_{A1} besagt hingegen, dass signifikante Mittelwerteunterschiede zwischen den Bewertungen der beiden Ratergruppen vorhanden sind und die Bewertungen nicht übereinstimmen. Mathematisch ausgedrückt ergeben sich somit folgende hypothetische Überlegungen: $\mathbf{H_1: ZR = SR; H_{A1}: ZR \neq SR}$. Der Test wird als zweiseitig beschrieben.

Für den Vergleich innerhalb einer Ratergruppe wird im Rahmen der Nullhypothese (H_2) davon ausgegangen, dass sich die Raterurteile einer Gruppe nicht unterscheiden, sondern gleich sind. Die Alternativhypothese H_{A2} dazu besagt, dass sie sich unterscheiden. Der Test ist zweiseitig. Mathematisch ausgedrückt ergeben sich somit für zwei beliebige Rater R1 und R2 die Hypothesen: $\mathbf{H_2: R1 = R2}$ **und** $\mathbf{H_{A2}: R1 \neq R2}$. Wie sie sich unterscheiden ist für die Fragestellung allerdings uninteressant, es ist lediglich relevant zu wissen, ob die Rater trotz vorhergegangener Schulung signifikante Unterschiede in der Bewertung zeigen.

2. Material und Methoden

2.1 Materialliste

Die Materialien waren für die Pilotuntersuchung sowie für die Hauptuntersuchung gleich. Sie unterschieden sich lediglich in ihrer Anzahl entsprechend der teilnehmenden Studierenden und der Anzahl an Stationen. Die Codierung „X“ in der folgenden Materialliste steht für die Anzahl der Kandidaten, also im Falle der Pilotuntersuchung für die Anzahl „9“, während sie im Falle der Hauptuntersuchung für die Anzahl „36“ steht. Die Codierung „Y“ steht für die Anzahl der Stationen, also für die Pilotuntersuchung fünf und für die Hauptuntersuchung sechs Stationen. Die Materialliste für die Station BLS wurde nur für die Hauptuntersuchung verwendet, da diese Station in der Pilotuntersuchung noch nicht konzipiert war und erst in der Hauptuntersuchung pilotiert wurde. Da die Pilotuntersuchung noch papierbasiert bewertet wurde, erhielten die Rater dafür Zettel und Stift, während die Rater der Hauptuntersuchung ein Tablet-PC (iPad) verwendeten.

Allgemein:

- Y Türschilder mit Aufgabenstellungen
- Y Uhren
- 8 Tablet-PCs, Serie iPad, Fa. Apple, Cupertino, Kalifornien, Vereinigte Staaten von Amerika
- 1 Trillerpfeife
- 12-16 Raumteiler, benutzbare Fläche 150 cm hoch, 125 cm breit, Gesamthöhe 200 cm
- 9 „sichere, digitale Speicherkarten“ (SD-Karten), 16 Gigabyte (GB), Fa. Hama, Monheim, Deutschland
- SD-Karten-Lesegerät, Fa. ISY, Typ ICR-2000 35 in 1 Alu Universal, Artikelnummer 1468788, Ingolstadt, Deutschland
- externe Festplatte, 1 Terabyte (TB), Fa. Toshiba, Typ HDTB330MK3CA Canvio Basics, Neuss, Deutschland
- 5 Tische
- Flipchart mit Laufplan
- 36 Sicherheitsnadeln
- 36 QR-Codes (auf Papier gedruckt), Matrix aus schwarzen und weißen Quadraten, die für spezifische Daten kodieren und scanbar sind; in diesem Fall für die Namen der Prüflinge
- Nitril® Best Gen® Handschuhe, 100 Stück, puderfrei, latexfrei, Größen small (S), medium (M) und large (L), Fa. Meditrade GmbH, Kiefersfelden, Deutschland

Videostationen:

- X Rezeptzettel (rot)
- 2X Terminkärtchen der UMG
- X Aufklärungsbögen für Zahnextraktion (Pelka und Farmand 2012), Herausgeber proCompliance, Firma Thieme Compliance GmbH, Verlagsort Erlangen, Deutschland
- Stoffhandtücher
- 2 Camcorder, SONY HDR- PJ410 B.CEN, schwarz
- 2 Hama Star 5, Dreibeinstativ für Camcorder, schwarz
- Orthopantomogramm mit Aufhellung am Zahn 26

Station Lokalanästhesie:

- Gentle Skin® Classic Handschuhe, puderfrei, unsteril, 100 Stück, Größen S, M, L, Fa. Meditrade GmbH, Kiefersfelden, Deutschland
- Suavel® Protect Mundschutz, 3-lagig mit elastischen Ohrschlaufen, Fa. Meditrade GmbH, Kiefersfelden, Deutschland
- Lokalanästhesiemodelle Oberkiefer (OK), Fa. Frasaco, Modellserie „AG-3 IB“, frasaco GmbH, Tettnang, Deutschland
- Modell Unterkiefer (UK), Fa. Frasaco, Modellserie „AG-3“, frasaco GmbH, Tettnang, Deutschland
- Grundbesteck (zahnärztliche Sonde und Spiegel, zahnärztliche Pinzette und Zungenspatel)
- X Sterican® Mix, Einwegkanülen, Material Kunststoff, Farbe nach der Europäische Norm der International Organization for Standardization (EN ISO) 6009 rot-violett, Außendurchmesser nach EN ISO 9626 1,4, Länge 40 mm, Größe in Gauge = 17, Fa. Braun Medical AG, Sempach, Schweiz
- X Sterican® für Dental Anästhesie, Einwegkanülen, Material Kunststoff, Farbe nach EN ISO 6009 mittel-grau, Außendurchmesser nach EN ISO 9626 0,4, Länge 22 mm, Größe in Gauge = 27, Fa. Braun Medical AG, Sempach, Schweiz
- X Injekt®, zweiteilige Einmalspritzen mit Lüer-Ansatz, Volumen 2 ml, 0,1 ml Skalenwert, Konus zentrisch, Fa. Braun Medical AG, Sempach, Schweiz
- X Brechampullen Ultracain DS, 1:200.000, 2 ml, 40 mg/ml/ 0,006 mg/ml Injektionslösung Articainhydrochlorid/ Epinephrinhydrochlorid, Fa. Sanofi Aventis Deutschland GmbH, Frankfurt, Deutschland
- ein steriles Operations-Tuch (OP-Tuch), Farbe grün
- Abwurfbehälter für scharfe Gegenstände

Station Zahnextraktion:

- Gentle Skin® Classic Handschuhe, puderfrei, unsteril, 100 Stück, Größen S, M, L, Fa. Meditrade GmbH, Kiefersfelden, Deutschland
- Suavel® Protect Mundschutz, 3-lagig mit elastischen Ohrschlaufen, Fa. Meditrade GmbH, Kiefersfelden, Deutschland
- Extraktionsmodelle OK und UK, Fa. Frasaco, Modellserie A-E-K, frasaco GmbH, Tettnang, Deutschland
- Extraktionswagen mit multiplem Sieb (Hebel nach Bein, Flohr, Barry, Extraktionszangen OK Frontzahn-, Prämolaren- (PM) und Molarenbereich (M), Extraktionszangen UK Frontzahn-, PM- und M-Bereich, Wurzelrest-Zange, Wundhaken nach Langenbeck, Wassmund und Middledorf, Wundhäkchen nach Gilles, zahnärztliche Sonde und Spiegel, zahnärztliche Pinzetten, scharfer Löffel nach Hemingway, Myrtenblattsonde, Hohlmeißelzange nach Lühr)
- zwei sterile OP-Tücher, Farbe grün
- X Askina® Schlinggaze-Tupfer, aus Verbandmull, pflaumengroß, Fa. Braun Medical AG, Sempach, Schweiz

Station Naht:

- Gentle Skin® Classic Handschuhe, puderfrei, unsteril, 100 Stück, Größen S, M, L, Fa. Meditrade GmbH, Kiefersfelden, Deutschland
- Suavel® Protect Mundschutz, 3-lagig mit elastischen Ohrschlaufen, Fa. Meditrade GmbH, Kiefersfelden, Deutschland
- Naht-Pad nach König, Fa. Erler-Zimmer, Lauf, Deutschland
- Nahtbesteck (chirurgische Pinzette, Nadelhalter nach Hegar-Mayo, Schere)
- Nadeln (gebogen), Fa. Johnson & Johnson Medical GmbH, Deutschland
- Seide traumatisch 3.0, Fa. Johnson & Johnson Medical GmbH, Neuss, Deutschland
- Seide atraumatisch 3.0, Fa. Johnson & Johnson Medical GmbH, Neuss, Deutschland
- Abwurfbehälter für scharfe Gegenstände

Station Basic Life Support:

- Decke
- Reanimationspuppe, Ambu® Man Modell I (Torso)
- 36 Actiomedic® MediSave Notfall-Beatmungstücher, Einmalprodukte, Fa. Sanismart GmbH, Waltrop, Deutschland

2.2 Methoden

2.2.1 Die OSCE in der zahnärztlichen Chirurgie der Universitätsmedizin Göttingen

Zu Beginn des Projektvorhabens „OSCE in der zahnärztlichen Chirurgie“ war es möglich, bei den seit Jahren in der Fakultät für Humanmedizin der Universitätsmedizin Göttingen stattfindenden OSCE-Prüfungen zu hospitieren, von den Erfahrungen der dort bereits etablierten Prüfungsform zu lernen und einen ersten Überblick über den Ablauf und die Struktur des Prüfungsformats zu erhalten. Im „Studentischen Trainingszentrum Ärztlicher Praxis und Simulation“, STÄPS, werden OSCE-Prüfungen schon seit Jahren mit hoher Akzeptanz der Studierenden durchgeführt (Simmenroth-Nayda et al. 2014).

Die Vorbereitungen für die erste Durchführung erforderten eine Prüfung des Projektvorhabens durch die Ethik-Kommission der Universitätsmedizin Göttingen, die am 28. Mai 2014 ohne weitergehende Bewertung genehmigt wurde. Im Hinblick auf die Datenschutz- und Persönlichkeitsrechte der Studierenden aufgrund der beiden geplanten Videostationen und des daraus entstehenden Bildmaterials wurde im Vorfeld ein Aufklärungsbogen entwickelt, der über den Verbleib und die Handhabung der Videodaten informierte und mit dem das Einverständnis von den Teilnehmern vor Prüfungsantritt schriftlich eingeholt wurde (siehe Anhang 7.1.1). Als Grundlage dieser schriftlichen Einverständniserklärung diente der für die OSCE-Prüfungen der Humanmediziner an der UMG verwendete Bogen (siehe Anhang 7.1.2).

Die Schauspielpatienten

Zur Durchführung der beiden kommunikativen OSCE-Stationen wurden erfahrene standardisierte Patienten/-innen (SP), auch Schauspielpatienten/- innen genannt, aus der SP-Kartei der humanmedizinischen Fakultät rekrutiert. Diese SP hatten bereits z. T. jahrelange Erfahrungen speziell auf dem Gebiet „OSCE-Prüfung“ gesammelt. Im Rahmen einer vorab stattfindenden Schulung wurden sie in ihre individuellen, standardisierten Rollen eingeführt. Sie erfuhren dabei sowohl die inhaltlichen Grundlagen ihrer Patientenrollen als auch die Charakterzüge und Gemütszustände, welche sie den Studierenden gegenüber darstellen sollten. Außerdem wurden sie mit den Räumlichkeiten der Stationen bekannt gemacht sowie mit den von den Prüflingen zu bewältigenden Aufgabenstellungen ihrer Station. Sie wurden mit gezielten Fragen und Äußerungen vertraut gemacht, um den Studierenden gegebenenfalls Hilfestellung leisten zu können und sie zurück auf den richtigen Pfad zu führen, sollten sie sich verlieren. Ein Beispiel dafür ist die Frage nach sportlicher Aktivität in den nächsten Tagen nach einer plastisch verschlossenen MAV.

Die Rater

Zur Bewertung der studentischen Performance wurden Ärzte und Zahnärzte aus der Klinik und Poliklinik für Mund, Kiefer- und Gesichtschirurgie (MKG) der UMG eingesetzt. Zwei der Stationen wurden im Hinblick auf die Schonung personeller Ressourcen als Videostationen geplant. Die Rater wurden den jeweiligen Stationen vorab zugeteilt und das Checklistenformat auf Basis der curricularen Vorlagen und Lernziele mithilfe eines Blueprints erstellt (Newble 2004). „Der Blueprint gewährleistet, dass alle Curriculumsziele geprüft werden und die entscheidenden Schlüsselprobleme angemessen repräsentiert sind“ (Nikendei und Jünger 2006, S. 2). Die Stationen und die Lernziele werden im weiteren Verlauf beschrieben. Die Inhalte der Bewertungsbögen (Raterbögen) wurden im Rahmen der Pilotstudie evaluiert und modifiziert, um anschließend in der Hauptuntersuchung verwendet zu werden.

Die Rater nahmen vorab an einer internen Schulung teil, die zur Erläuterung des Ablaufplans und der Aufgabenbesprechung diente. Sie erhielten dabei Einsicht in zu Schulungszwecken erstellte Videos, welche die Aufgabenstellungen der drei praktischen zahnärztlich-chirurgischen Stationen darstellten. Sie wurden mit den einzelnen Items konfrontiert und konnten sich ein Gesamtbild über die einzelnen Prüfungsstationen machen, so wie es von anderen Autoren vorgeschlagen wird (Kiehl et al. 2014). Außerdem erhielten sie individuelle Instruktionen zum Thema Bewertungsformat und der verwendeten Tablet-Software.

Die Prüflinge erhielten im Vorfeld eine Beschreibung der im Rahmen der Prüfung zu bewältigenden Stationen, um sich gezielt auf die Prüfungsinhalte vorbereiten zu können. Während die Bewertung der Pilotuntersuchung noch papierbasiert durchgeführt wurde, so bewerteten die Rater bereits im ersten formativen OSCE ein halbes Jahr später (im Oktober 2014) Tablet-basiert.

Das Bewertungsformat

Die Rater bewerteten mithilfe der Checklisten und vergaben zusätzlich eine Globalnote. Codierungen für die Checklisten-Items ermöglichten eine zeitnahe statistische Auswertung (Kiehl et al. 2014). Das Gesamtergebnis setzte sich zu 80% aus den Ergebnissen der Checklistenbewertung und zu 20% aus den Ergebnissen der Globalbewertung zusammen. Die Vergabe der Globalnote zielt hierbei darauf ab „interaktive Kompetenzebenen wie Empathie" (Nikendei und Jünger 2006, S. 2) besser miteinzubeziehen, während die Checklistenbewertung vermehrt geschlossene Fragen beinhaltet. Aus der Literatur wird deutlich, dass eine Kombination der beiden Bewertungen empfohlen wird. "Checklisten sind inhaltsspezifischer, während die globale Beurteilung ein breiteres Spektrum an Fertigkeiten erfasst" (Chenot und Ehrhardt 2003, S. 5), wodurch die Testgüte erhöht werden kann. Hierbei wird "bei Stationen mit vorrangig technischen Fertigkeiten eine Checkliste als adäquates Evaluationsinstrument, bei der Prüfung von kommunikativen Kompetenzen hingegen die Verwendung eines globalen Ratings" (Nikendei und Jünger 2006, S. 2) empfohlen (Newble 2004; Reznick

et al. 1996). Auch in anderen Fachrichtungen stellt sich die Vergabe einer Globalnote gleichermaßen und teilweise sogar als besser geeignet dar (Read et al. 2015; Regehr et al. 1998).

2.2.2 Ablauf der Voruntersuchung: Pilot-OSCE

Die Pilotierung der Stationen der ersten formativen OSCE-Prüfung in der zahnärztlichen Chirurgie der Universitätsmedizin Göttingen fand am 18.07.2014 im Studentischen Innovations- und Trainingszentrum Zahnmedizin (SINUZ) statt. Neun freiwillige Studierende des neunten Semesters nahmen an der Studie teil. Sie durchliefen nacheinander fünf Stationen mit praktischen Aufgaben aus dem täglichen Arbeitsfeld der zahnärztlichen Chirurgie: (1) Präoperatives Aufklärungs- und Einwilligungsgespräch vor Zahnextraktion, der Patient/ die Patientin (SP) wird über die Vorgehensweise und möglicherweise auftretende Komplikationen aufgeklärt, und die schriftliche Einwilligung wird eingeholt; (2) postoperatives Patientengespräch nach aktuell plastisch verschlossener Mund-Antrum-Verbindung; (3) Lokalanästhesie am Simulationsmodell; (4) Zahnextraktion am Simulationsmodell; (5) Naht am Simulator (Naht Pad nach König). Die Stationen werden im Folgenden noch genauer beschrieben. Ein Parcours-Plan war am Eingang des SINUZ für die Prüflinge einsehbar. Sie wurden vorab instruiert in Behandlungskleidung mit weißem Kittel zu erscheinen und eine Schutzbrille mitzubringen. Alle Anwesenden wurden während des Prüfungstages mit Getränken und Essen verköstigt.

Die Studierenden durchliefen den Parcours in Kleingruppen von fünf Studierenden, sodass jeder Station zeitgleich ein Prüfling zugeteilt werden konnte. Es verblieben ihnen zwei Minuten Zeit, um die jeweilige Aufgabenstellung der Station zu lesen und zu verinnerlichen, bevor ein lauter Pfiff als akustisches Startsignal den Beginn der Prüfungszeit von fünf Minuten definierte. Nach den fünf Minuten wies ein erneuter Pfiff auf einen Stationswechsel hin und die Teilnehmer rotierten zur nächsten Station. Dort verblieben ihnen erneut zwei Minuten zum Lesen der Aufgabenstellung, die Wechsel- und Lesezeit betrug also zwei Minuten. Die Prüfung galt als beendet, sobald alle Stationen durchlaufen worden waren.

Die Bewertung der studentischen Leistungen erfolgte mit zwei unterschiedlichen Verfahren. An den beiden Kommunikationsstationen waren Videokameras aufgebaut, die das Gespräch zwischen Student und SP aufzeichneten. Die Bedienung der Kameras übernahmen die Studenten und Studentinnen bei Betreten und Verlassen des Raumes eigenverantwortlich. Die Auswertung durch zahnärztliche Rater erfolgte zeitnah versetzt. An den anderen drei Stationen war jeweils ein approbierter Zahnarzt/ eine approbierte Zahnärztin vor Ort und bewertete papierbasiert.

Station 1 – präoperatives Aufklärungsgespräch

An Station 1 trafen die Studierenden auf eine Schauspielpatientin namens Frau G., die langjährige Patientin in der Praxis ist. Frau G. ist 57 Jahre alt und bis auf leicht erhöhten und gut eingestellten Blutdruck allgemeinanamnestisch unauffällig, sie zeigt eine gute Mundhygiene und hat keine wirtschaftlichen Probleme. Die Aufgabenstellung sieht vor, dass der Patientin bereits am letzten Termin mitgeteilt wurde, dass ein subgingivaler Füllungsrand am Zahn 26 kariös ist, der Zahn nicht erhaltungswürdig ist und somit extrahiert werden muss. Die Patientin wollte über diese Nachricht noch einmal schlafen, stellt sich erneut in der Praxis vor und soll von dem Prüfling über die Extraktion aufgeklärt werden, um ihre Einwilligung zu geben.

Neben den fachlichen Aspekten, über die die Studierenden den Patienten aufklären sollen, spielen hier psychosoziale Aspekte eine Rolle. Die SP sind während der Vorbereitungsphase individuell instruiert worden gewisse Zweifel und Sorgen in Bezug auf einen derartigen Eingriff vorzubringen. Hier soll der Patient beruhigt und ihm die Angst vor einem derartigen Routineeingriff genommen werden und zusätzlich die wichtigsten Komplikationen fachlich korrekt dargelegt bekommen. Dabei sollte der in der UMG verwendete und den Studierenden während der Prüfung vorliegende Aufklärungsbogen zur Zahnextraktion Verwendung finden (Pelka und Farmand 2012).

Station 2 – postoperatives Aufklärungsgespräch

In der Station 2 saß Frau H. (SP), der gemäß Aufgabenstellung soeben der Zahn 17 entfernt worden war. Bei der Extraktion des oberen Molaren 17 war es zu einer Mund-Antrum-Verbindung gekommen, die nach erfolgreicher Extraktion des Zahnes sofort plastisch gedeckt worden war. Der Student/ die Studentin soll der Patientin nun kurz den durchgeführten Eingriff näher erläutern und in patientengerechten Worten, ohne zahnärztliches Fachvokabular verständlich darstellen, worauf aufgrund der aufgetretenen MAV in den nächsten Tagen zu achten ist. Falls der Prüfling vergisst, relevante Aspekte des postoperativen Verhaltens anzusprechen, so gibt die SP, wie im Rahmen der Schulung instruiert, den Studierenden mit gezielten Fragen Hilfestellung.

Station 3 – Lokalanästhesie

In der Station 3 sollten die Prüflinge am Phantommodell den Zahn 26 zur Extraktion anästhesieren. Alle benötigten Materialien lagen dafür bereit. Die Studierenden sollen während der gesamten Prüfungszeit hygienisch einwandfreies Verhalten demonstrieren und den Zahn vor dem Hintergrund der bevorstehenden Extraktion mit der richtigen Technik und den richtigen Volumina, wie in den bereits erwähnten Lehrveranstaltungen vermittelt, korrekt anästhesieren.

Station 4 – Extraktion

Station 4 dient der Überprüfung der Fertigkeit der Studierenden zur Zahnextraktion. Hierbei sollte am Situationsmodell der Zahn 17 extrahiert werden, die LA ist bereits erfolgt. Alle benötigten Materialien liegen für die Studierenden bereit. Es ist erneut wichtig zu zeigen, dass der Umgang mit den bereitliegenden Instrumentarien bekannt ist, diese sicher ausgewählt und unter hygienisch korrekten Bedingungen verwendet werden können.

Station 5 – Naht

Station 5 überprüft die Fähigkeiten der Studenten/ Studentinnen eine suffiziente Nahtversorgung durchzuführen. Hierzu dient das Nahtpad nach König als Simulationsmodell, welches die natürliche Beschaffenheit der Haut nachempfindet. Ziel ist es dabei, eine traumatische Einzelknopfnaht mit einem Handknoten und eine atraumatische Einzelknopfnaht mit einem Nadelhalterknoten durchzuführen. Alle benötigten Materialien liegen dafür bereit. Die Studierenden sollen hier abermals hygienisch und fachlich korrektes Vorgehen demonstrieren.

Nach Absolvierung der OSCE wurden die Studierenden gebeten eine anonymisierte Abschlussevaluation auszufüllen. Die Evaluation umfasste eine Selbsteinschätzung der jeweiligen Stationen anhand der Globalnote, eine Einschätzung von der erfolgreichen/ nicht erfolgreichen Absolvierung der Prüfung und eine Bewertung des Prüfungsformats im Allgemeinen (Schiekirka et al. 2012). Dabei wurden die Studierenden befragt, inwieweit sie die OSCE Prüfung als angemessen beschreiben würden, für wie sinnvoll sie die Kommunikationsstationen mit standardisierten Patienten halten und wie ihr Urteil im Vergleich zu Multiple Choice Klausuren im Fach zahnärztliche Chirurgie ausfällt. Auch der generelle Wunsch nach strukturierter Lehre bezüglich Patientenkommunikation für das Zahnmedizinstudium wird hier erfragt, siehe Anhang 7.1.3.

Die ärztlichen Rater bewerteten die Raterbögen und den Ablauf sowie die Organisation der OSCE, wodurch es möglich war, eine zeitnahe Überarbeitung der Checklisten durchzuführen. Auch die statistischen Ergebnisse hatten Einfluss auf die Überarbeitung der Bögen und die Aufgabenstellung, sowie die Logistik. Die überarbeiteten Raterbögen wurden daraufhin in der folgenden OSCE im Oktober 2014 verwendet.

2.2.3 Ablauf der Hauptuntersuchung: formative OSCE Oktober 2014

Nach entsprechender Überarbeitung der Logistik, der Listen und Inhalte der Raterbögen fand die erste formative OSCE daraufhin am 17.10.2014 als Semesterabschlussprüfung im Fach der zahnärztlichen Chirurgie im SINUZ der Zahnklinik Göttingen statt und ist seitdem integrativer Bestandteil

des dritten klinischen Semesters. Die 36 Studenten und Studentinnen, davon acht männlich und 28 weiblich, durchliefen am Tag der Prüfung in Kleingruppen die gleichen, bereits pilotierten Stationen. An der Aufgabenstellung änderte sich hierbei teilweise die zu behandelnden Regionen. Die Prüfungszeit betrug für alle Teilnehmenden dabei insgesamt 42 Minuten, wovon eine Minute pro Station Wechsel- und Lesezeit und sechs Minuten pro Station Prüfungszeit waren. Der Parcours-Plan war wie bereits bei dem OSCE-Pilot am Eingang des SINUZ aufgestellt und für die Prüflinge einsehbar. Alle Teilnehmenden, Studierenden, standardisierte Patienten und Rater wurden während des Tages mit Essen und Getränken verköstigt. Die Studierenden waren lediglich angewiesen in Behandlungskleidung mit weißem Kittel zu erscheinen und eine Schutzbrille mitzubringen. Der in der IT erstellte QR-Code zum Einscannen mittels der Tablets befestigte jeder Prüfling mit Sicherheitsnadeln an der Rückseite des Kittels.

Zusätzlich zu den oben genannten fünf Stationen (1) präoperatives Aufklärungsgespräch, (2) postoperatives Aufklärungsgespräch, (3) Lokalanästhesie, (4) Zahnextraktion und (5) Nahtversorgung kam eine weitere, sechste Station zur Pilotierung hinzu: (6) Die Reanimation einer Trainingspuppe (BLS Torso) als lebensrettende Sofortmaßnahme in einer Notfallsituation.

Station 6 – Basic Life Support

An dieser Station wird angenommen, dass der Prüfling als behandelnde(r) Zahnarzt/Zahnärztin den Patienten nach einer Leitungsanästhesie im Behandlungszimmer kurz allein gelassen hat und der Patient bei Betreten des Raums reglos neben dem Behandlungsstuhl liegt. Hier wird der Prüfling aufgefordert seine notfallmedizinischen Fertigkeiten zu demonstrieren und die einwandfreie Durchführung lebensrettender Sofortmaßnahmen (Basic Life Support – BLS) an der Reanimationspuppe unter Beweis zu stellen.

Im Unterschied zur Pilot-OSCE wurden alle sechs Prüfungsstationen sowohl von studentischen als auch von zahnärztlichen bzw. ärztlichen Ratern gleichzeitig und unabhängig voneinander bewertet. Beide Ratergruppen bewerteten Tablet-basiert auf Grundlage der bereits pilotierten, evaluierten und überarbeiteten Raterbögen. Die BLS Station stellte eine Ausnahme dar, an der die Checklisten-Items erst in diesem Durchlauf pilotiert wurden. Die beiden Gesprächsstationen wurden per Videoaufnahme dokumentiert und zeitversetzt ausgewertet, während an den anderen vier Stationen die jeweiligen Rater von studentischer und ärztlicher Seite anwesend waren und eine primäre Bewertung durchführten. Auch die Teilnehmer und Teilnehmerinnen gaben nach der OSCE eine Selbsteinschätzung ihrer Leistungen ab (Kiehl et al. 2014). Der dazugehörige Bogen ist im Anhang unter 7.1.3 zu finden.

Als studentische Rater konnten die freiwilligen Prüflinge der Pilot-OSCE eingesetzt werden, zu diesem Zeitpunkt Studierende im 10. Semester. Die zahnärztlichen und ärztlichen Rater aus der

MKG sowie dem Zentrum für Anästhesiologe, Rettungs- und Intensivmedizin (ZARI) und die studentischen Rater nahmen vorab an einer ausführlichen Raterschulung teil. In dieser Raterschulung erhielten sie Einblick in zu Schulungszwecken erstellte Videos und wurden individuell zum Thema Tablet-Benutzung und Verhaltensweisen während der Prüfung instruiert. Die Videos zeigten die Prüfungsinhalte der jeweiligen Stationen und orientierten sich dabei an den Lehrinhalten der Operationskurse der UMG. In den Videos demonstrieren Lehrkörper der MKG, wie die entsprechenden Aufgaben praktisch durchgeführt werden sollen, so zum Beispiel die Lokalanästhesie oder die Extraktion eines Molaren. Die Vorgehensweisen werden praktisch demonstriert und zusätzlich erklärt. Die Videos haben dabei eine Länge von ca. fünf Minuten. Die zentralen Punkte dieser Demonstration finden sich inhaltlich in den Checklisten-Items wieder. Zusätzlich erhielten die Rater eine Einweisung in die Handhabung der iPads und Bewertungs-App, um technisch verschuldeten Problemen vorzubeugen. Darüber hinaus wurden sie über die Abläufe der Prüfung, wie beispielsweise die Bedeutung der Pfeiftöne informiert sowie über die Räumlichkeiten, das Procedere an ihrer Station und ihre Funktion als Rater aufgeklärt. Sie erhielten Anweisungen keine Hilfestellungen während der Prüfung zu geben und die Performance der Studierenden nicht zu kommentieren. Ziel der Raterschulung ist es, eine Standardisierung der Rater und somit der Prüfungsbedingungen für die Studierenden zu gewährleisten (Nikendei und Jünger 2006).

Im Anschluss an ihre Arbeit evaluierten die Rater die Tablet-Benutzung. Die Evaluation der verwendeten Tablet-Software wurde von beiden Ratergruppen ausgeführt. Sie enthielt zum einen Fragen zur Bewertung der Funktionalität der Software, also zur Effizienz der Funktionen in der Applikation und zu deren Anwendung. Außerdem wurde die Software-Ergonomie in Bezug auf die Verständlichkeit der verwendeten Symbole und Bezeichnungen, die Erkennbarkeit und Unterscheidbarkeit der verwendeten Farben und Darstellungen und die Führung durch die Applikation evaluiert. Auch die Fehlerrobustheit wurde bewertet und die Möglichkeit zur Korrektur von Eingabefehlern und die Verständlichkeit von Fehlermeldungen erfragt. Die 14 geschlossenen Fragen des Fragebogens konnten in den sechs Abstufungen „trifft voll zu“, „trifft weitgehend zu“, „trifft überwiegend zu“, „trifft selten zu“, „trifft sehr selten zu“ und „trifft gar nicht zu“ bewertet werden. Der zugehörige Bogen ist im Anhang 7.1.4 zu finden.

2.2.4 Die elektronische Datenauswertung

Die Etablierung der Tablet-Software sowie die Verarbeitung erworbenen Daten erfolgte zunächst im Institut für Medizinische Informatik (IT) der Universitätsmedizin Göttingen. Zu diesem Zweck kam die App tOSCE zum Einsatz, welche auf den für die Bewertung verwendeten iPads installiert wurde. Die App erfasste zunächst die aus der OSCE-Prüfung erhaltenen Datensätze. Außerdem erstellt die IT QR-Codes, welche jedem Studierenden ausgehändigt wurden und das Einscannen mittels iPad ermöglichten. Die Grundauszählung und Dokumentation erfolgte mit Microsoft Excel 2013 ebenfalls in der IT. Die anschließende Auswertung der Datensätze wurde mithilfe des Umbrella Consortium for Assessment Networt (UCAN) - Modulsystems der Universität Heidelberg vorgenommen. Dieses basiert auf der ItemManagementSystem (IMS) Software. Die Checklisten wurden ebenfalls mithilfe der UCAN erstellt. Die ausgewerteten Daten konnten vom UCAN-Server zurück nach Göttingen überspielt und via SPSS statistics 22 statistisch ausgewertet werden.

Im Rahmen der statistischen Auswertung wurde sowohl eine Item-Analyse der einzelnen Items jeder Station erstellt als auch die deskriptive Statistik und die Verteilungseigenschaften der Funktionen, die sich aus den einzelnen Stationen ergeben, erhoben. Die Interraterreliabilität für die zahnärztlichen und die studentischen Rater wurde mithilfe eines t-Tests für abhängige Stichproben und zugehöriger Bravais-Pearson-Korrelation sowie mit einer Intra-Klassen-Korrelation vorgenommen.

2.2.5 Die statistischen Methoden

Die statistische Auswertung der Ergebnisse der Itemanalyse, also der Einzelkomponenten (Items) der Checklisten für die jeweiligen Stationen, erfolgte anhand von Codierungen. Die Items wurden in Punktwerte übersetzt. Jedes Item konnte mit „gut“ bzw. „trifft zu“, „mittel gut“ und „schlecht“ bzw. „trifft nicht zu“ beurteilt werden. Dabei bekommen „gut“ und „trifft zu“ einen Punktwert von 1, „mittel“ einen von 0,5 und „schlecht“ bzw. „trifft nicht zu“ einen von 0. Die Itemanalyse gab Aufschluss über die Konsistenz der Checklisten (Cronbachs Alpha) und Eignung der Items (Item-Schwierigkeit und Trennschärfe).

Die Pilot-OSCE

Die Items zur Bewertung der fünf Stationen der Pilot-OSCE wurden für die Stichprobengröße N=9 analysiert. Dazu wurden jeweils die Mittelwerte, die zugehörige Standardabweichung und die Trennschärfen erhoben, ebenso wie der Gesamtmittelwert für die Station. Die genauen Daten der Itemanalyse werden im Folgenden noch beschrieben. Der wörtliche Inhalt der einzelnen Items der

Stationen wurde im Hinblick auf die Durchführbarkeit der folgenden OSCE-Prüfungen codiert. Der Checklisten-Gesamtwert wird dazu als Mittelwert über die gesamten Items einer Stationsauswertung berechnet und lässt aufgrund der Codierung somit gleich eine prozentuale Einschätzung erkennbar werden. Die Noten erhielten ebenfalls eine Codierung: 1=1, 2=0, 3=0,75, 4=0,25 und 5=0. Cronbachs Alpha wurde für die gesamte Station berechnet, sowie zusätzlich als Gesamtwert über die Station unter dem jeweiligen Ausschluss der einzelnen Items.

Die Parameter der deskriptiven Statistik, die Verteilungseigenschaften Schiefe, Kurtosis, Minimum und Maximum wurden für die Checklisten-Gesamtwerte der einzelnen Stationen sowie für die umcodierten Noten und das Gesamtergebnis erhoben. Dieses setzte sich für die Pilot-OSCE aus 80% Checklistenbewertung und 20% Note zusammensetzt. Die Stichprobengröße beträgt dabei für die Pilot-OSCE ebenfalls N=9.

Die Hauptuntersuchung

Zur statistischen Analyse wurden die Mittelwerte (M), die Standardabweichung (SD) und die Trennschärfe sowohl für die einzelnen Checklisten-Items als auch für die jeweiligen Stationen insgesamt errechnet. Der Mittelwert gibt dabei die Schwierigkeit des Items und der Station an. Zusätzlich wurde Cronbachs Alpha für die einzelnen Stationen errechnet sowie der Wert für Cronbachs Alpha, der sich durch die Eliminierung des jeweiligen Items für die Gesamt-Station ergeben würde ("α if deleted"). Diese Daten wurden zum einen für die Bewertung der ärztlichen Rater, zum anderen für die Bewertung der studentischen Rater erhoben. Zusätzlich wurde die Korrelation von Checklisten-Mittelwert und Gesamteindruck der jeweiligen Stationen für Zahnarzt und studentischen Rater getrennt ermittelt. Die genauen statistischen Messwerte der einzelnen Items sind den beigefügten Tabellen 23-28 im Anhang zu entnehmen, siehe 7.3.

Zusätzlich zur Itemanalyse wird die deskriptive Statistik für die einzelnen Stationen erhoben. Die Stichprobengröße beträgt dabei 36 (N=36). Anhand der deskriptiven Statistik lassen sich Aussagen über die Verteilungseigenschaften der Merkmale „Checklisten-Mittelwerte (chl)“, „Gesamteindruck (Note)“ und „Kombination aus beiden (gesamt)“ treffen. Die Berechnung erfolgte sowohl für die zahnärztlichen Rater (A) und die studentischen Rater (S) unabhängig voneinander als auch für die Selbsteinschätzung der Studierenden. Bei der Selbsteinschätzung variierte die Stichprobe N aufgrund von Enthaltungen zwischen 31 und 34. Die in der Globalbewertung zu erreichenden Punkte waren wie folgt codiert: Die Punktzahl 0 entspricht einer Note 5 (mangelhaft); 0,5 = 4 (ausreichend); 1 = 3 (befriedigend); 1,5 = 2 (gut) und 2=1 (sehr gut). Die zu erreichende Maximalpunktzahl entspricht mit 2 Punkten somit einer Note von 1 (Minimum der Verteilung) und umgedreht die Minimalpunktzahl 0 einer Note 5 (Maximum der Verteilung). Die Globalnote und die entsprechende Punktzahl sind folglich entgegengerichtet codiert, sodass ein „niedriger“ Notenwert und eine hohe Punktzahl

jeweils einer guten Leistung entsprechen. Checklisten-Mittelwerte und Globalbewertung ergeben mit einer Gewichtung von 70% zu 30% das Gesamtergebnis. Bei den Checklisten-Mittelwerten entsprechen hohe Werte ebenfalls einer guten Leistung. Das bedeutet, dass sie mit der Note an sich entgegengerichtet codiert sind, mithilfe der entsprechenden Punktzahl jedoch wieder gleichgerichtet.

Die Interrater-Reliabilität wurde mittels Intra-Klassen-Korrelation (ICC) und Produkt-Moment-Korrelation berechnet. Die Erläuterungen zu diesen beiden Korrelationen befinden sich im Einleitungsteil. Für die Mittelwertevergleiche zwischen den Fremdurteilen der zahnärztlichen und der studentischen Rater wurden die statistischen Kennungen für t-Verteilungen erhoben: die Differenz der Mittelwerte und die zugehörige Standardabweichung, der T-Wert und die Anzahl an Freiheitsgraden mit dem dazugehörigen p-Wert. Außerdem wurde der Korrelationskoeffizient r berechnet und ebenfalls der dazugehörige p-Wert. Eine Wiederholung der statistischen Kennungen ist auch der Tabelle 1 in der Einleitung zu entnehmen.

3. Ergebnisse

3.1 Ergebnisse – Pilot-OSCE

3.1.1 Itemanalyse der Pilot-OSCE-Stationen

Der Ausdruck „Mittelwert M“ bezeichnet im Folgenden den Mittelwert über alle Studierenden für das jeweilige Checklisten-Item oder die gesamte Station, errechnet mithilfe der im Material- und Methodenteil erklärten Codierungen. Der „Item-Mittelwert“ oder die „Item-Schwierigkeit“ bezieht sich dabei auf ein bestimmtes Checklisten-Item. Er kann gleichzeitig als Interpretation der Schwierigkeit verstanden werden, wonach ein hoher Wert für den Mittelwert eine geringe Schwierigkeit, also ein „leichtes“ Checklisten-Item bezeichnet. Der „Checklisten-Mittelwert“ oder die „Checklisten-Schwierigkeit“ hingegen setzt sich aus der Gänze der Item-Mittelwerte zusammen und definiert die Schwierigkeit der gesamten Station. Mit „Stations-Mittelwert“ oder „Stations-Schwierigkeit“ wird das sich aus den Checklistenmittelwerten und Globalnote zusammensetzendes Endergebnis für die jeweilige Station bezeichnet. Die Gewichtung dieser Zusammensetzung wird in dem Teil zur jeweiligen Untersuchung genannt.

Diese Bezeichnungen werden auch auf die anderen statistischen Kennungen übertragen: die „Item-Trennschärfe“ (für jedes Item) wird also von der „Checklisten-Trennschärfe“ (gemittelt über alle Items) unterschieden. Gleiches gilt für die Standardabweichung SD. Die Bezeichnung „Gesamteindruck“ wird hier als Synonym für die Bewertung mithilfe einer Globalnote verwendet.

In den Tabellen 2-6 der folgenden Abschnitte sind die Mittelwerte (M) über alle teilnehmenden Studierenden mit zugehöriger Standardabweichung (SD) und Trennschärfe sowie „Alpha if deleted“ dargestellt. Gelb unterlegt sind die auffälligen Werte für die Trennschärfen unter 0,3. Sie sind dann auffällig, wenn die zugehörige Schwierigkeit hoch ist. Dies wird im folgenden Abschnitt erläutert. Die zugehörigen Schwierigkeiten sind in den Tabellen mit der Farbe „blau“ unterlegt. Die letzte Zeile stellt die Korrelation zwischen Note (Globalnote) und Skalenmittelwert (Checklisten-Mittelwert) dar.

3.1.1.1 Station „Aufklärung und Einwilligung Zahnextraktion“

Wie in Tabelle 2 zu sehen, schwanken die zwölf Item-Mittelwerte zwischen 0,31 (SD 0,46) und 0,94 (SD 0,18) und ergeben einen Checklisten-Mittelwert von 0,69 (SD 040). Dieser Wert besagt also, dass 69% der Studierenden die Gesamtheit der Checklisten-Items richtig gelöst haben.

Die Item-Trennschärfen variieren zwischen -0,44 und 0,84. Das Item „verständliche Wortwahl“ ergibt eine negative Trennschärfe von -0,44. Dieses Item zeigt mit einer Schwierigkeit von 0,81, dass 81% der Studierenden das Item lösten, die verbleibenden 19% allerdings von denjenigen Studierenden schlecht gelöst wurden, die in der restlichen Station gut abschnitten.

Die Korrelation zwischen der Bewertung mittels Checklisten-Mittelwerten und der Bewertung mittels Globalnote ist mit 0,96 hoch und signifikant. Das negative Vorzeichen entsteht aufgrund der entgegen gerichteten Codierung von Note und Skalenmittelwert. Je größer der Wert für die Note wird, desto kleiner wird der entsprechende Checklisten-Mittelwert (Skalenmittelwert). Ein großer Wert für die Note und ein kleiner Wert für die Checkliste entsprechen in dieser Form einer schlechten Leistung, und umgekehrt.

Tabelle 2: Die Itemanalyse der Station „Aufklärung und Einwilligung Zahnextraktion“

Item	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item weggelassen
Item 1	0,75	0,38	0,18	0,81
Item 2	0,94	0,18	0,42	0,80
Item 3	0,81	0,37	-0,44	0,85
Item 4	0,81	0,37	0,63	0,77
Item 5	0,81	0,26	0,73	0,77
Item 6	0,63	0,52	0,55	0,78
Item 7	0,63	0,52	0,81	0,74
Item 8	0,56	0,42	0,46	0,79
Item 9	0,75	0,46	0,84	0,74
Item 10	0,88	0,35	0,62	0,77
Item 11	0,38	0,52	0,55	0,78
Item 12	0,31	0,46	0,29	0,80
Mittelwerte	**0,69**	**0,40**	**0,47**	**0,78**
Cronbachs Alpha	0,80			
Korrelation	-0,96			

3.1.1.2 Station „Lokalanästhesie“

Die 13 Item-Mittelwerte ergeben an dieser Station Werte zwischen 0,00 (SD 0,00) und 1,00 (SD 0,00) und eine Checklisten-Schwierigkeit von 0,64 (SD 0,37), siehe Tabelle 3.

Tabelle 3: Die Itemanalyse der Station „Lokalanästhesie“

Item	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item weggelassen
Item 1	0,89	0,33	-0,46	0,52
Item 2	1,00	0,00	0,00	0,41
Item 3	0,78	0,44	0,17	0,38
Item 4	0,56	0,53	0,62	0,17
Item 5	0,67	0,50	-0,25	0,52
Item 6	0,78	0,44	0,17	0,38
Item 7	0,67	0,50	0,14	0,39
Item 8	1,00	0,00	0,00	0,41
Item 9	0,67	0,50	0,29	0,33
Item 10	0,00	0,00	0,00	0,41
Item 11	0,44	0,53	0,19	0,37
Item 12	0,33	0,50	0,45	0,26
Item 13	0,56	0,53	0,29	0,33
Mittelwerte	**0,64**	**0,37**	**0,13**	**0,38**
Cronbachs Alpha	0,41			
Korrelation	-0,85			

Die Mehrzahl der Items hat eine Trennschärfe von unter 0,3, zwei von ihnen befinden sich im negativen Bereich. Cronbachs Alpha ergibt für die Station einen Wert von 0,41.

Die Korrelation zwischen Note und Skalenmittelwert ist auch an dieser Station mit -0,85 hoch und signifikant.

3.1.1.3 Station „Zahnextraktion“

Diese Station zeigt mit Ausnahme von zwei der 14 Items hohe Item-Schwierigkeiten zwischen 0,72 (SD 0,44) und 1,00 (SD 0,00). Die beiden Items, die niedrigere Schwierigkeiten zeigen, stellen inhaltlich die Überprüfung der korrekt sitzenden Anästhesie zu Beginn der Extraktion (Item 5), sowie der nach Zahnextraktion möglicherweise aufgetretenen MAV (Item 13) dar (siehe Tabelle 4). Der Checklisten-Mittelwert beträgt 0,80 (SD 0,29). Die Trennschärfen von 0,00 sind Resultat der Schwierigkeit von 1,00. Die Checklisten-Trennschärfe liegt bei 0,27. Drei weitere Items erhielten negative Trennschärfen. Cronbachs Alpha zeigt für die Stationen einen Wert von 0,63. Würde man das Kriterium „Prüft LA mit Spiegel und Sonde“ (Item 5) herausnehmen, so erhöht es sich auf 0,71. Die Korrelation zwischen Note und Skalenmittelwert liegt bei -0,55.

Tabelle 4: Die Itemanalyse der Station „Zahnextraktion“

Item	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item weggelassen
Item 1	1,00	0,00	0,00	0,63
Item 2	1,00	0,00	0,00	0,63
Item 3	0,78	0,26	0,29	0,60
Item 4	0,94	0,17	0,05	0,63
Item 5	0,33	0,50	-0,15	0,71
Item 6	0,83	0,25	0,59	0,57
Item 7	0,89	0,33	-0,31	0,69
Item 8	0,83	0,35	-0,25	0,69
Item 9	0,72	0,44	0,61	0,53
Item 10	0,83	0,35	0,89	0,48
Item 11	0,78	0,44	0,79	0,48
Item 12	0,94	0,17	0,05	0,63
Item 13	0,44	0,53	0,41	0,57
Item 14	0,89	0,33	0,75	0,52
Mittelwerte	**0,80**	**0,29**	**0,27**	**0,60**
Cronbachs Alpha	0,63			
Korrelation	-0,55			

3.1.1.4 Station „Naht“

Die Nahtstation zeigt einen Checklisten-Mittelwert von 0,77 (SD 0,29), also eine Erfolgsquote von 77%. Die einzelnen Werte sind in Tabelle 5 zu finden. Die Item-Mittelwerte schwanken zwischen 0,39 (SD 0,33) und 1,00 (SD 0,00). Item 5, 11 und 12 zeigen niedrige Mittelwerte (M = 0,39; M = 0,44; M = 0,50). Die Checklisten-Trennschärfe liegt bei 0,3. Die Trennschärfen von drei Items (Item 4, 8 und 9) zeigen negative Werte zwischen -0,17 und -0,04.

Tabelle 5: Die Itemanalyse der Station „Naht“

Item	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item weggelassen
Item 1	0,89	0,33	0,83	0,57
Item 2	1,00	0,00	0,00	0,68
Item 3	0,72	0,44	0,41	0,64
Item 4	0,83	0,25	-0,13	0,71
Item 5	0,39	0,33	0,28	0,66
Item 6	0,89	0,22	0,77	0,61
Item 7	1,00	0,00	0,00	0,68
Item 8	0,89	0,33	-0,17	0,72
Item 9	0,89	0,22	-0,04	0,69
Item 10	0,89	0,33	0,83	0,57
Item 11	0,44	0,46	0,00	0,72
Item 12	0,50	0,35	0,63	0,60
Item 13	0,67	0,50	0,55	0,61
Mittelwerte	**0,77**	**0,29**	**0,30**	**0,65**
Cronbachs Alpha	0,68			
Korrelation	-0,94			

Cronbachs Alpha ergibt für die Station einen Wert von 0,68. Würde das Item mit der niedrigsten Trennschärfe (Item 8 mit -0,17) weggelassen, so erhöht sich Cronbachs Alpha auf 0,72. Die Korrelation zwischen Note und Skalenmittelwert ist mit -0,94 hoch und signifikant.

3.1.1.5 Station „Aufklärung post OP“

Bei Item-Mittelwerten zwischen 0,00 (SD 0,00) und 0,83 (SD 0,35) ist für diese Station der niedrigste Checklisten-Mittelwert mit 0,54 (SD 0,36) zu finden. Dieser bezeichnet folglich die höchste Schwierigkeit.

Tabelle 6: Die Itemanalyse der Station „Aufklärung post OP“

Item	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item weggelassen
Item 1	0,33	0,50	0,00	0,74
Item 2	0,56	0,39	0,42	0,67
Item 3	0,61	0,42	0,19	0,70
Item 4	0,83	0,35	0,37	0,68
Item 5	0,61	0,42	0,57	0,64
Item 6	0,78	0,36	0,57	0,65
Item 7	0,61	0,42	0,65	0,63
Item 8	0,44	0,30	0,34	0,68
Item 9	0,44	0,30	0,76	0,63
Item 10	0,72	0,26	0,04	0,71
Item 11	0,44	0,46	0,16	0,71
Item 12	0,00	0,00	0,00	0,70
Item 13	0,61	0,49	0,29	0,69
Mittelwerte	**0,54**	**0,36**	**0,33**	**0,68**
Cronbachs Alpha	0,70			
Korrelation Note und Skalenmittelwert	-0,81			

Das Item „fragt nach Krankmeldung“ (Item 12) erhält eine Schwierigkeit von 0,00 (SD 0,00). Dies bedeutet, dass kein Studierender dem Patienten/ der Patientin eine Krankschreibung offeriert. Die Stations-Trennschärfe liegt bei 0,33. Cronbachs Alpha zeigt einen Wert von 0,70. Die Korrelation zwischen dem Skalenmittelwert und der Note ist mit -0,81 hoch und signifikant.

3.1.2 Deskriptive Statistik und Verteilungseigenschaften

3.1.2.1 Checklisten-Mittelwerte

Die Item-Schwierigkeiten der einzelnen Stationen wurden bereits dargestellt. Als kurze Wiederholung ist zu sagen, dass sich die Schiefe zwischen -0,5 und + 0,5 und die Kurtosis zwischen -1 und +1 als annährend normalverteilt bezeichnen lässt.

Tabelle 7: Checklisten-Gesamtwerte

Dargestellt sind hier die Stichprobengröße N, das Minimum (Min) und Maximum (Max), der Checklisten-Mittelwert der jeweiligen Station mit zugehöriger Standardabweichung SD sowie die Parameter der Verteilungseigenschaften Schiefe und Kurtosis. Gelb markiert sind hier die Werte für Schiefe und Kurtosis, die nicht mehr als normalverteilt gelten.

Checklisten-Gesamtwert (Mittelwert über alle Items einer Station)	N	Min	Max	Mittelwert	SD	Schiefe	Kurtosis
„Aufklärung und Einwilligung Zahnextraktion“	9	0,29	0,96	0,66	0,24	-0,39	-1,38
„Lokalanästhesie“	9	0,46	0,92	0,64	0,15	0,86	-0,09
„Zahnextraktion“	9	0,50	0,93	0,80	0,14	-1,49	2,04
„Naht“	9	0,42	0,88	0,77	0,15	-1,84	3,85
„Aufklärung post OP“	9	0,31	0,81	0,54	0,18	0,35	-1,13

Die Verteilung für die Station „Aufklärung und Einwilligung Zahnextraktion“ zeigt eine Schiefe von -0,39 und eine Kurtosis von -1,38. Die Kurtosis ist somit außerhalb des normalverteilten Bereichs und zeigt eine platykurtische Form. Dazu ist jedoch zu sagen, dass die Breite der Form bei einer Stichprobengröße von N=9 nur bedingt aussagekräftig ist. Aufgrund der geringen Größe ist die Verteilung nicht resistent gegenüber Ausreißern und wird durch diese stark beeinflusst. Sobald mehrere

Studierende das gleiche Ergebnis erzielen hat dies bereits Auswirkungen auf die Verteilungseigenschaften, was bei größerer Stichprobenanzahl kaum Auswirkungen hat. Die Häufigkeitsverteilungen für die Stationen sind zur Verdeutlichung im Anhang 7.2 zu finden.

Die Verteilungseigenschaften der Station „Lokalanästhesie“ zeigen eine stark linkssteile (Sch = 0,86) Form mit einem Minimum von 0,46 und einem Maximum von 0,92. Die Stationen „Zahnextraktion“ und „Naht“ zeigen beide stark rechtssteile und leptokurtische Eigenschaften. Die Verteilungseigenschaften der Checklisten-Mittelwerte der Station „Aufklärung post OP“ zeigen mit einer Kurtosis von -1,13 eine breitgipflige Form.

3.1.2.2 Noten nach Umcodierung

Die Verteilungseigenschaften für die Noten nach Umcodierung ergeben für die Leistungen der Studierenden Mittelwerte zwischen 0,47 (SD 0,20) und 0,72 (SD 0,26), siehe Tabelle 8. Die Station „Aufklärung und Einwilligung Zahnextraktion“ ergibt eine Schiefe und eine Kurtosis von jeweils -0,55. Dies beschreibt eine rechtssteile Verteilung.

Tabelle 8: Umcodierte Noten

Dargestellt sind hier die Stichprobengröße N, das Minimum (Min) und Maximum (Max), der Checklisten-Mittelwert der jeweiligen Station mit zugehöriger Standardabweichung SD sowie die Parameter der Verteilungseigenschaften Schiefe und Kurtosis. Gelb markiert sind die Werte für Schiefe und Kurtosis, die nicht mehr als normalverteilt gelten.

Umcodierte Noten	N	Min	Max	Mittelwert	SD	Schiefe	Kurtosis
„Aufklärung und Einwilligung Zahnextraktion“	9	0,25	1,00	0,72	0,26	-0,55	-0,55
„Lokalanästhesie“	9	0,25	0,75	0,47	0,20	0,22	-1,04
„Zahnextraktion“	9	0,00	1,00	0,47	0,34	-0,13	-0,78
„Naht“	9	0,25	0,75	0,64	0,18	-1,50	1,47
„Aufklärung post OP“	9	0,25	1,00	0,56	0,21	1,17	2,43

Für die Station „Lokalanästhesie“ zeigt die Kurtosis mit -1,04 eine platykurtisch Form. An der Station „Zahnextraktion“ erhalten die Studierenden Noten zwischen 0,0 und 1,0. Die Note 5, also umcodiert der Punktwert 0,0 wurde hier durch Extrahieren des falschen Zahnes erreicht. Die Station ist die einzige, an der dieser Notenwert vergeben wurde. Die Eigenschaften der Verteilung befinden sich im annährend normalverteilten Bereich.

Die Station „Naht“ ergibt eine stark rechtssteile (Sch = -1,50) und schmalgipfige (Ex = 1,47) Verteilung.

Auch die Station „Aufklärung post OP“ zeigt keine Merkmale einer Normalverteilung. Die entstandene Verteilung ist stark linkssteil mit einer Schiefe von 1,17 und besitzt stark leptokurtische Eigenschaften mit einer Kurtosis von 2,43.

3.1.2.3 Das Gesamtergebnis

Die Gesamtergebnisse werden in Prozent angegeben und stellen dar, wie die Studierenden insgesamt abgeschnitten haben. Sie setzen sich für die Pilotuntersuchung zu 80% aus den Checklisten-Mittelwerten über alle teilnehmenden Studierenden und zu 20% aus der Note zusammen. Die Station „Naht“ erhält mit 74,32% die höchste Prozentzahl, die Station „Lokalanästhesie“ mit 60,73% die niedrigste, siehe Tabelle 9.

Das Gesamtergebnis für die Pilot-OSCE, bestehend aus den Gesamtergebnissen aller Stationen, ergibt 65,95% (SD 13,21). Die entsprechende Verteilung entspricht keiner Normalverteilung, sondern stellt eine rechtssteile, leptokurtische Verteilung dar. Die entsprechenden Werte für Schiefe und Kurtosis sind Tabelle 9 zu entnehmen.

Die Station „Aufklärung und Einwilligung Zahnextraktion“ zeigt rechtssteile, stark platykurtische Verteilungseigenschaften. Für die Station „Lokalanästhesie“ ist die Verteilung linkssteil und platykurtisch. Sowohl die Station „Zahnextraktion“, als auch die Station „Naht“ stellen stark rechtssteile und stark schmalgipflige Funktionen dar. Die Werte für die Station „Aufklärung post OP“ zeigen annährend normalverteilte Eigenschaften.

Tabelle 9: Gesamtergebnisse aus Checkliste und Globalnote

Dargestellt sind hier die Stichprobengröße N, das Minimum (Min) und Maximum (Max), der Checklisten-Mittelwert der jeweiligen Station mit zugehöriger Standardabweichung SD sowie die Parameter der Verteilungseigenschaften Schiefe und Kurtosis. Gelb markiert sind die Werte für Schiefe und Kurtosis, die nicht mehr als normalverteilt gelten.

Gesamtwert (80% Checkliste, 20% Globalnote)	N	Min	Max	Mittelwert	SD	Schiefe	Kurtosis
„Aufklärung und Einwilligung Zahnextraktion“	9	28,33	96,67	66,97	23,96	-0,44	-1,23
„Lokalanästhesie“	9	41,92	88,85	60,73	15,39	0,72	-0,36
„Zahnextraktion“	9	40,00	94,29	73,57	15,90	-1,19	1,78
„Naht“	9	38,85	85,77	74,32	15,28	-1,81	3,52
„Aufklärung post OP“	9	29,62	81,54	54,19	17,74	0,48	-0,69
Gesamt	9	38,74	82,59	65,95	13,21	-0,90	1,52

3.2 Ergebnisse Hauptstudie

3.2.1 Itemanalyse

Im Gegensatz zur Pilot-OSCE beträgt die Stichprobengröße - Anzahl der Studierenden - der Hauptstudie 36 (N=36). Es kommt eine sechste Station zur Pilotierung hinzu: die Station „Basic Life Support“. Die Items der Stationen wurden anhand der Erkenntnisse aus der Pilot-OSCE überarbeitet, präzisiert und erweitert. Die genauen Modifikationen der Items nach der Pilot-OSCE werden in der Diskussion ausführlicher beschrieben. Die Prüfungsbedingungen blieben unverändert. Die Bewertung durch die zahnärztlichen und ärztlichen Rater erhält im Folgenden den Buchstaben „A“, während die Bewertung durch die studentischen Rater den Buchstaben „S“ erhält.

3.2.1.1 Station „Aufklärung und Einwilligung Zahnextraktion“

Die Checklisten-Mittelwerte sind in Tabelle 10 dargestellt. Die Checklisten-Mittelwerte, die Checklisten-Standardabweichungen und die Checklisten-Trennschärfen ergeben bei den zahnärztlichen und studentischen Ratern ähnliche Ergebnisse. Auch Cronbachs Alpha zeigt ein ähnliches Ergebnis. Wäre der Gesamteindruck ein Item der Skala, so würde sich der Reliabilitätskoeffizient α auf 0,778 (A) und 0,782 (S) erhöhen. Die Tabellen zur Itemanalyse der einzelnen Stationen sind im Anhang unter 7.3.1 zu finden (Tabellen 23-28). Die Korrelation zwischen Checkliste und Globalnote ist mit 0,901 (A) und 0,861 (S) hoch und signifikant, siehe Tabelle 23.

Tabelle 10: Bewertung der Stationen durch die ärztlichen (A) und studentischen (S) Rater
Dargestellt sind hier die Checklisten-Mittelwerte, Checklisten-Trennschärfen und Cronbachs Alpha Koeffizienten der einzelnen Stationen sowie sie entsprechenden über die Gesamt-OSCE gemittelten Werte dargestellt. Grün unterlegte Felder heben die guten Kennwerte hervor, rote Felder die problematischen Werte.

Station	Checklisten-Mittelwerte (A)	Checkliste-Trenn-schärfe (A)	Cron-bachs Alpha (A)	Checkliste-Mittelwerte (S)	Checklisten-Trenn-schärfe (S)	Cron-bachs Alpha (S)
„Aufklärung und Einwilligung Zahnextraktion“	0,661	0,360	0,725	0,660	0,384	0,744
„postoperative Aufklärung“	0,734	0,469	0,811	0,687	0,295	0,631
„Lokalanästhesie“	0,677	0,165	0,481	0,686	0,106	0,328
„Zahnextraktion“	0,711	0,190	0,530	0,737	0,242	0,635
„Naht“	0,744	0,255	0,621	0,709	0,252	0,630
„BLS“	0,671	0,099	0,280	0,719	0,098	0,307
Gesamt-OSCE	0,700	0,256	0,575	0,700	0,230	0,546

3.2.1.2 Station „postoperative Aufklärung“

Die Item-Schwierigkeiten der zwölf Items dieser Station ergeben für die Bewertung durch die ärztlichen Rater einen Checklisten-Mittelwert von 0,734 mit einer SD von 0,329, siehe Tabelle 10. Die Station zeigt mit 0,811 den höchsten Wert für Cronbachs Alpha bei schwankenden Trennschärfen

zwischen 0,181 und 0,694. Mit dem Gesamteindruck als Item der Skala würde sich Cronbachs Alpha sogar auf 0,837 erhöhen. Die Korrelation zwischen Checklisten-Mittelwert und Gesamteindruck beträgt 0,874 (A) und 0,831 (S). Die Ergebnisse korrelieren hoch und signifikant ($p < 0,001$), siehe Tabelle 24.

Die Item-Mittelwerte, der Checklisten-Mittelwert und Cronbachs Alpha ergeben für die Bewertung durch die studentischen Rater niedrigere Werte als die Vergleichswerte der ärztlichen Rater. Der Wert für Cronbachs Alpha liegt bei 0,694.

3.2.1.3 Station „Lokalanästhesie“

Diese Station beinhaltet 14 Items, die zur Bewertung herangezogen werden. Die Item-Schwierigkeiten nach Bewertung durch die Zahnärzte ergeben einen Checklisten-Mittelwert von 0,677 (SD 0,404), siehe Tabelle 10. Cronbachs Alpha beträgt 0,481. Die Checklisten-Trennschärfe liegt bei 0,165. Der Großteil der Item-Trennschärfen liegt unter 0,3. Würde der Gesamteindruck hinzugezogen werden, so erhöhte sich Cronbachs Alpha auf 0,560.

Auch die Bewertungen durch die studentischen Rater zeigen ähnliche Ergebnisse. Der Checklisten-Mittelwert ergibt 0,686 (SD 0,394) und Cronbachs Alpha 0,328. Die Trennschärfen liegen zwischen -0,234 und 0,485, wovon drei Werte im negativen Bereich liegen.

Die Korrelation von Checklisten-Mittelwert und Gesamteindruck ist mit 0,827 (A) und 0,802 (S) hoch und signifikant, siehe Tabelle 25.

3.2.1.4 Station „Zahnextraktion“

Die Item-Schwierigkeiten der 16 zur Bewertung herangezogenen Items ergeben an dieser Station, bewertet durch die Zahnärzte, einen Checklisten-Mittelwert von 0,711 (SD 0,320), siehe Tabelle 10. Cronbachs Alpha beträgt 0,53, die Checklisten-Trennschärfe 0,190. Unter Einbeziehung des Gesamteindrucks erhöht sich Cronbachs Alpha auf 0,594.

Der Checklisten-Mittelwert aus der Bewertung durch die studentischen Rater beträgt 0,737 (SD 0,319). Cronbachs Alpha ist hier mit 0,635 höher als der Vergleichswert und verbessert sich durch den Gesamteindruck sogar auf 0,695.

Die Korrelation zeigt für die Zahnärzte einen Korrelationskoeffizienten von 0,741 und für die Studenten 0,831. Somit korrelieren beide hoch und signifikant ($p < 0,001$), siehe Tabelle 26.

3.2.1.5 Station „Naht“

Die Checklisten-Mittelwerte für die Station zeigen ähnliche Werte: 0,744 (A) und 0,709 (S). Cronbachs Alpha fällt mit 0,621 (A) und 0,603 (S) ebenfalls ähnlich aus. Unter Berücksichtigung des Gesamteindrucks ergibt sich für die zahnärztlichen Rater ein Alpha-Koeffizient von 0,675 und für die studentischen Rater 0,691, siehe Tabelle 10. Die Checklisten-Trennschärfe beträgt 0,255 (A). Der Großteil der Werte liegt unter 0,3. Bei der Bewertung durch die studentischen Rater liegt die Checklisten-Trennschärfe bei 0,252. Die Korrelation zwischen den Checklisten-Mittelwerten und dem Gesamteindruck ist mit 0,768 (A) und 0,836 (S) hoch und signifikant ($p < 0,001$), siehe Tabelle 27.

3.2.1.6 Station „Basic Life Support“

Diese Station zeigt für die Bewertung durch die ärztlichen Rater einen Checklisten-Mittelwert von 0,671 (SD 0,376), siehe Tabelle 10. Cronbachs Alpha ergibt mit 0,280 den niedrigsten Wert unter allen Stationen. Unter Einflussnahme des Gesamteindrucks erhöht sich Cronbachs Alpha auf 0,428. Auch die Checklisten-Trennschärfe ist mit 0,099 unterhalb des Referenzbereichs von 0,3. Zwei der Trennschärfen befinden sich mit -0,001 und -0,256 im negativen Bereich.

Die Bewertung durch die studentischen Rater zeigt einen Checklisten-Mittelwert von 0,719 (SD 0,323). Die Trennschärfen der Einzelitems liegen mehrheitlich unter 0,3. Durch den Gesamteindruck verbessert sich Cronbachs Alpha von 0,307 auf 0,431. Der Checklisten-Mittelwert und der Gesamteindruck korrelieren mit 0,827 (A) und 0,829 (S) hoch und signifikant ($p < 0,001$), siehe Tabelle 28.

3.2.1.7 Gesamt-OSCE

Die Itemanalyse für die Gesamt-OSCE zeigt nach der Bewertung durch die ärztlichen Rater eine Schwierigkeit von 0,700 (M) und eine Trennschärfe von 0,256. Cronbachs Alpha liegt bei 0,575 (M), siehe Tabelle 10. Für die Bewertung durch die studentischen Rater liegen die Schwierigkeit der Gesamt-OSCE ebenfalls bei 0,700 und die Trennschärfe bei 0,230. Cronbachs Alpha ergibt einen Wert von 0,546.

Die Korrelation zwischen den Checklisten-Mittelwerten und dem Gesamteindruck war über alle Stationen hoch und signifikant. Dies beschreibt einen positiven, linearen Zusammenhang zwischen den beiden Kriterien.

3.2.2 Deskriptive Statistik und Verteilungseigenschaften

3.2.2.1 Verteilungseigenschaften am Beispiel der Station „Aufklärung und Einwilligung Zahnextraktion“ (AEZ)

Die Verteilungseigenschaften werden hier für die Station „Aufklärung und Einwilligung Zahnextraktion“ einmal erläutert. Die entsprechenden Werte der anderen Stationen sind im Anhang 7.3 zu finden, werden hier aber nicht im Detail erläutert.

Wie bereits in der Itemanalyse erwähnt beträgt der Checklisten-Mittelwert (chl) für die zahnärztliche Bewertung 0,66 (SD 0,18). Schiefe und Kurtosis sind hierbei im annähernd normalverteilten Bereich. Für die Note liegt ein Mittelwert von 2,56 (SD 1,13) vor. Das Minimum liegt bei 1, das Maximum bei 5. Insgesamt ergibt sich für die Station ein Mittelwert von 0,65 (SD 0,21) und eine breitgipflige Verteilung aufgrund einer Kurtosis von -1,04.

Tabelle 11: Deskriptive Statistik der Station AEZ

Die Abkürzungen in den Spalten sind wie folgt zu verstehen: Chl = Checklisten-Mittelwerte; Note = Globalnote; A = zahnärztliche/ärztliche Rater; S = studentische Rater. In den Zeilen sind hier die Stichprobengröße N, der Checklisten-Mittelwert mit zugehöriger Standardabweichung SD sowie die Parameter der Verteilungseigenschaften Schiefe und Kurtosis und das Minimum (Min) und Maximum (Max) der Funktion dargestellt. Rot markiert sind die Werte für Schiefe und Kurtosis, die nicht mehr als normalverteilt gelten.

	AEZ_chl_A	AEZ_Note_A	AEZ_A_gesamt	AEZ_chl_S	AEZ_Note_S	AEZ_S_gesamt	AEZ_Note_Selbst
N	36	36	36	36	36	36	33
Mittelwert	0,66	2,56	0,65	0,66	2,39	0,66	2,45
SD	0,18	1,13	0,21	0,18	0,84	0,18	0,75
Schiefe	0,21	0,10	0,17	-0,33	0,06	-0,25	**0,86**
Kurtosis	-0,93	-0,92	**-1,04**	-0,22	-0,45	-0,22	0,07
Minimum	0,380	1	0,260	0,290	1	0,280	1
Maximum	1	5	1	1	4	1	4

Die Bewertung durch die studentischen Rater zeigt ähnliche Ergebnisse für die jeweiligen Mittelwerte und das Gesamtergebnis und annähernd normalverteilte Eigenschaften.

Die Selbsteinschätzung der Studierenden (N = 33) in Form der Globalnote siedelt sich mit 2,45 (SD 0,75) genau in der Mitte zwischen studentischer und ärztlicher Bewertung an. Keiner der

Prüflinge bewertet sich selbst mit einer schlechteren Note als „4“, die Mehrheit schätzt ihre Leistungen mit der Note „2“ ein. Die zugehörige Verteilung ist mit einer Schiefe von 0,86 linkssteil. Die Kurtosis zeigt die Eigenschaften einer Normalverteilung.

Zusammenfassend lässt sich somit eine Bilanz über die Stationen ziehen. Für die Bewertung des Gesamt-Konstruktes durch die beiden verschiedenen Ratergruppen Ärzte und Studenten ergaben sich annährend normalverteilte Funktionseigenschaften. Dies bezieht sich auf alle Stationen für die Gesamtbewertung aus Checklisten-Mittelwerten und Note.

Die Selbsteinschätzung der Studierenden zeigt Auffälligkeiten in ihren Verteilungseigenschaften. Die Gesamtbewertung dieser Selbsteinschätzung ergibt für die Schiefe linkssteile Eigenschaften. Diese Eigenschaft resultiert daraus, dass die Studierenden ihre Prüfungsleistungen durchschnittlich besser bewerteten als die zahnärztlichen und ärztlichen Rater. Eine Ausnahme ist dabei die Station „Naht“.

3.2.2.2 Verteilungseigenschaften der Gesamt-OSCE

Die Verteilungseigenschaften werden ebenfalls für die Gesamt-OSCE, also über alle Stationen, erfasst.

Die Bewertung der studentischen Rater ergibt in Bezug auf den Mittelwert für die Gesamt-OSCE vergleichbare Werte (M = 0,67). Die Funktion ist normalverteilt. Das Maximum liegt dabei bei 1, das Minimum bei 0,5. Die Gesamtnote fällt mit einem Mittelwert von 2,56 leicht besser aus als die Bewertung durch die Ärzte. Die zugehörigen Eigenschaften ähneln denen einer Normalverteilung. Der Checklisten-Mittelwert beträgt 0,7 und zeigt somit keinen Unterschied zur Bewertung der Ärzte. Die Funktion ist normalverteilt.

Über die Gesamtergebnisse der ärztlichen Rater (A) lässt sich sagen, dass sie annährend normalverteilte Eigenschaften zeigen. Der Mittelwert der Gesamt-OSCE liegt bei 0,66 (siehe Tabelle 12). Der Minimalwert beträgt 0,460, der Maximalwert 1. Für die Gesamt-OSCE vergeben die zahnärztlichen Rater Globalnoten zwischen „2“ und „4“, bei einem Mittelwert von 2,76.

Tabelle 12: Deskriptive Statistik für die Gesamt-OSCE

In den Spalten werden hier die Gesamt-Checklisten-Mittelwerte (gesamt_chl) über alle Stationen, die Gesamt-Globalbewertung (gesamt_Note) sowie das Gesamtergebnis, kombiniert aus beiden Merkmalen für jeweils die Bewertung durch die zahnärztlichen und die studentischen Rater getrennt voneinander (gesamt_OSCE) dargestellt. In den Zeilen sind die Stichprobengröße N, der Checklisten-Mittelwert mit zugehöriger Standardabweichung SD sowie die Parameter der Verteilungseigenschaften Schiefe und Kurtosis und das Minimum (Min) und Maximum (Max) der Funktionen angegeben. Rot markiert sind die Werte für Schiefe und Kurtosis, die nicht mehr als normalverteilt gelten.

	gesamt_ chl_A	gesamt_ Note_A	gesamt_ OSCE_A	gesamt_ chl_S	gesamt_ Note_S	gesamt_ OSCE_S	gesamt_ Note_Selbst
N	36	36	36	36	36	36	34
Mittelwert	0,70	2,76	0,66	0,70	2,56	0,67	2,53
SD	0,08	0,49	0,09	0,07	0,47	0,09	0,45
Schiefe	-0,05	-0,08	0,09	-0,17	0,17	-0,19	0,61
Kurtosis	0,15	0,19	0,15	-0,17	-0,87	-0,50	0,46
Minimum	0,540	2	0,460	0,550	2	0,500	2
Maximum	1	4	1	1	4	1	4

Die Selbsteinschätzung über die Gesamt-OSCE mit N = 34 ergibt einen Mittelwert von 2,53. Dieser ist im Vergleich zu der Bewertung durch die beiden Ratergruppen niedriger. Das bedeutet, dass die Studierenden ihre eigenen Leistungen besser einschätzten, als diese tatsächlich durch die verschiedenen Rater bewertet wurden. Diese Erkenntnis ist aus der Literatur bereits bekannt (Schiekirka et al. 2014). Die Verteilung kann nicht mehr als normalverteilt beschrieben werden, da die Schiefe mit 0,61 linkssteile Eigenschaften zeigt. Keiner der Studierenden vergab die Note „5“ oder die Note „1“.

3.2.3 Vergleich der Mittelwerte mit t-Test für abhängige Stichproben

3.2.3.1 Auswertung der t-Verteilung der Bewertung durch zahnärztliche und studentische Rater

Im Rahmen der t-Tests wurden für jedes Bewertungskriterium jeder Station - also die Checklisten-Mittelwerte (chl), die Note (Note) und die Kombination aus Checkliste und Note (gesamt) - die entsprechenden Verteilungseigenschaften erhoben. Daraus lässt sich das Gesamtergebnis für die OSCE-Prüfung darstellen. Die Mittelwertevergleiche der beiden Ratergruppen sind Abbildung 2 zu entnehmen.

Tabelle 13: Vergleich der zahnärztlichen und der studentischen Rater

In den Spalten sind die Stichprobengröße N, Schwierigkeit/ Mittelwert M, die Standardabweichung SD, der T-Wert T, die Anzahl an Freiheitsgraden df, der p-Wert und der Korrelationskoeffizient r dargestellt. p (2-seitig) von r bezieht sich auf die Signifikanz der Korrelation r, p (2-seitig) auf die Signifikanz der Mittelwerte-Differenzen. In den Zeilen befinden sich die jeweiligen Ergebnisse der Stationen unter den bereits erklärten Bedingungen. Die Farbe „grün" weist auf ein signifikantes Ergebnis p < 0,05 hin, während die Farbe „rot" nicht signifikante Ergebnisse p > 0,2 unterstreicht. Die Farbe „orange" markiert Werte, die im Indifferenzbereich liegen: 0,05 < p < 0,2.

		Zahnarzt		studentischer Rater		Differenz						
	N	M	SD	M	SD	M	SD	T	df	p (2-seitig)	R	p (2-seitig) von r
AEZ_Note	36	2,56	1,13	2,39	0,84	0,17	1,00	1,00	35	0,324	0,52	0,001
POA_Note	36	2,33	0,83	2,61	0,77	-0,28	0,94	-1,77	35	0,086	0,30	0,075
LAE_Note	36	2,89	0,89	2,86	0,80	0,03	0,70	0,24	35	0,812	0,66	0,000
ZE_Note	36	3,08	0,84	2,72	1,06	0,36	0,72	3,00	35	0,005	0,73	0,000
Naht_Note	36	2,89	0,71	2,50	0,91	0,39	0,77	3,05	35	0,004	0,58	0,000
BLS_Note	36	2,81	0,95	2,25	0,73	0,56	0,61	5,49	35	0,000	0,77	0,000
Gesamt-Note	36	2,76	0,49	2,56	0,47	0,20	0,27	4,57	35	0,000	0,85	0,000
AEZ_chl	36	0,66	0,18	0,66	0,18	0,00	0,13	0,05	35	0,959	0,73	0,000
POA_chl	36	0,73	0,19	0,69	0,13	0,05	0,15	1,88	35	0,069	0,64	0,000
LAE_chl	36	0,68	0,15	0,69	0,13	-0,01	0,10	-0,62	35	0,543	0,78	0,000
ZE_chl	36	0,71	0,13	0,74	0,14	-0,03	0,09	-1,68	35	0,102	0,76	0,000
Naht_chl	36	0,74	0,13	0,71	0,15	0,03	0,10	1,96	35	0,058	0,74	0,000
BLS_chl	36	0,67	0,14	0,72	0,12	-0,05	0,11	-2,75	35	0,009	0,67	0,000
Gesamt-Chl	36	0,70	0,08	0,70	0,07	0,00	0,04	-0,07	35	0,942	0,88	0,000
AEZ_gesamt	36	0,65	0,21	0,66	0,18	-0,01	0,16	-0,45	35	0,659	0,68	0,000
POA_gesamt	36	0,71	0,19	0,66	0,15	0,05	0,17	1,93	35	0,062	0,54	0,001
LAE_gesamt	36	0,63	0,17	0,64	0,15	-0,01	0,10	-0,53	35	0,597	0,80	0,000
ZE_gesamt	36	0,64	0,14	0,69	0,17	-0,05	0,11	-2,56	35	0,015	0,78	0,000
Naht_gesamt	36	0,68	0,14	0,68	0,17	-0,01	0,12	-0,26	35	0,794	0,72	0,000
BLS_gesamt	36	0,63	0,16	0,71	0,14	-0,08	0,10	-4,40	35	0,000	0,77	0,000
gesamt_OSCE	36	0,66	0,09	0,67	0,09	-0,02	0,04	-2,34	35	0,025	0,90	0,000

Die Stichprobengröße N beträgt für jede der Verteilungen 36, die Anzahl an Freiheitsgraden df folglich 35 (N-1). Die Tabelle 13 zeigt die jeweiligen Mittelwerte und zugehörige Standardabweichung. Für die Mittelwertedifferenz wird der Mittelwert der studentischen Bewertung von dem Mittelwert

der zahnärztlichen Bewertung subtrahiert. Die Gesamt-Note setzt sich aus den Noten aller sechs Stationen zusammen. Das gleiche gilt für die Checklisten-Mittelwerte (Gesamt-chl). Das Gesamtergebnis (_gesamt) setzt sich zu 70% aus den Checklisten-Mittelwerten und zu 30% aus der Note zusammen. Aus den Parametern ergeben sich somit für jedes Kriterium sieben Wertegruppen, insgesamt also 21 Wertegruppen.

Die Mittelwertedifferenzen schwanken zwischen -0,28 und 0,56. Der Betrag der Werte gibt die Abweichungen voneinander an. 15 der 21 Mittelwertevergleiche ergeben eine Differenz von 0,00 (SD 0,13) bis 0,08 (0,10). 60% dieser Werte sind im Bereich von 0,00 (SD 0,13) bis 0,03 (SD 0,10). Die übrigen Mittelwertevergleiche zeigen Differenzen von 0,17 (SD 1,00) bis 0,56 (SD 0,61). Diese Mittelwertedifferenzen betreffen die Globalnote, was aus der Skalierung resultiert. Im Vergleich zu den Checklisten-Mittelwerten kann die Note Werte zwischen 1,0 und 5,0 annehmen.

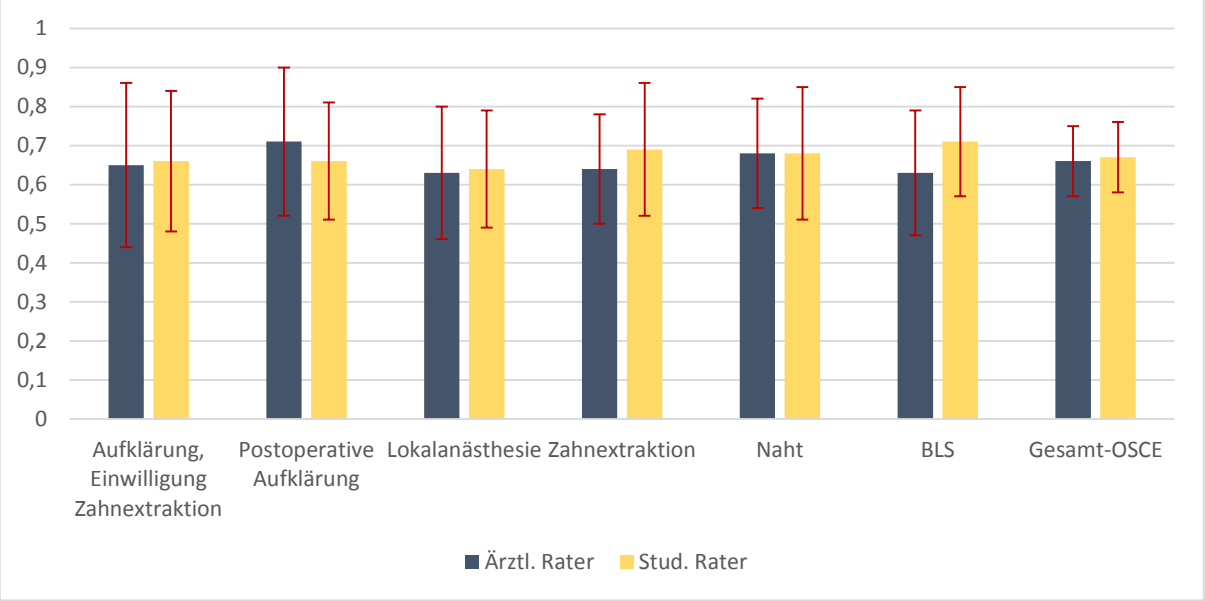

Abbildung 2: Mittelwertevergleich der ärztlichen und studentischen Rater

In Abbildung 2 werden die Mittelwertevergleiche in Form von Balkendiagrammen dargestellt. Auf der x-Achse sind die Gesamtbewertungen der sechs Stationen, bestehend aus Checklisten-Mittelwerten und Globalnoten, und das Gesamtergebnis der OSCE aufgetragen, während auf der y-Achse die Mittelwerte (Schwierigkeiten) zu erkennen sind. Möglich sind dabei Werte zwischen 0,0 und 1,0. Die Farbe „blau" entspricht der Bewertung durch die ärztlichen Rater, die Farbe „gelb" der Bewertung durch die studentischen Rater. Die roten Balken stellen die zugehörigen Standardabweichungen dar.

Der Vergleich zwischen der Bewertung der zahnärztlichen und der studentischen Rater zeigt die in Tabelle 13 dargestellten p-Werte: Acht der Werte liegen mit p-Werten zwischen 0,324 und 0,959 im

nicht-signifikanten Bereich und fünf der p-Werte liegen im Indifferenzbereich. Sie ergeben Werte zwischen 0,058 und 0,102 und können somit in beide Richtungen interpretiert werden. Ebenfalls fünf Ergebnisse liegen im Bereich zwischen 0,004 und 0,025 und deuten darauf hin, dass das Ergebnis signifikant ist. Drei der errechneten p-Werte liegen bei 0,00. Die Interpretation dieser Ergebnisse erfolgt im Diskussionsteil.

Der Vergleich der beiden Ratergruppen zeigt durchgehend positive Korrelationskoeffizienten zwischen 0,30 und 0,90. Der r-Wert für die Bewertung durch die Vergabe einer Note für die Station „postoperative Aufklärung“ ist mit 0,30 am kleinsten und deutlich kleiner als alle anderen Werte. Der entsprechende p-Wert zeigt ein nicht-signifikantes Ergebnis, da p = 0,075. Alle übrigen p-Werte liegen im signifikanten Bereich und machen somit deutlich, dass die hohe Korrelation zwischen den Mittelwerten der beiden Ratergruppen kein Zufall ist, sondern dass eine Systematik zugrunde liegt. Es liegen sehr hohe und signifikante Zusammenhänge zwischen beiden Fremdurteilen vor. Liegen signifikante Mittelwerteunterschiede vor, so liegt dies oft an der hohen Korrelation, da in die Formel zur Berechnung der Mittelwerteunterschiede mittels t-Test für abhängige Stichproben die Korrelation mit eingeht. Aus diesem Grund wurde zusätzlich eine Intra-Klassen-Korrelation über die Werte erhoben, die genauere Aussagen über die Signifikanz zulässt.

3.2.3.2 Vergleiche zwischen Fremd- und Selbsteinschätzung (Globalnoten)

Vergleich Bewertung der Zahnärzte – Selbsteinschätzung der Studierenden

Die Stichprobengröße dieses t-Tests variiert zwischen N = 31 und N = 34 und somit folglich auch die Anzahl der Freiheitsgrade df der t-Verteilung zwischen N = 30 und N = 33. Die zugehörigen Werte zeigt die Tabelle 14. Zur Berechnung der Mittelwertedifferenzen der Vergleichsgruppen werden die Noten der Selbstbeurteilung von denen durch Zahnärzte vergebenen Noten subtrahiert. Es entstehen dadurch ausschließlich positive oder keine Mittelwertedifferenzen mit Werten zwischen 0,00 und 0,81. Dies zeigt, dass die Studierenden sich an jeder Station besser bewerteten, als sie durch die Zahnärzte bewertet wurden.

Die Werte besagen, dass ein sehr enger Zusammenhang zwischen der Bewertung von Zahnärzten und der Selbsteinschätzung durch die Studierenden besteht. Der Zusammenhang ist positiv und linear. Eine höhere Bewertung durch die zahnärztlichen Rater entspricht also auch einer höheren Bewertung durch die Studierenden selbst, und umgekehrt.

Tabelle 14: Vergleich Bewertung der Zahnärzte – Selbsteinschätzung der Studierenden

N = Stichprobengröße; M = Schwierigkeit/ Mittelwert; SD = Standardabweichung; T = T-Wert; df = Anzahl an Freiheitsgraden; p = p-Wert, r = Korrelationskoeffizient. Die Farbe „grün" weist auf ein signifikantes Ergebnis p < 0,05 hin, während die Farbe „rot" nicht signifikante Ergebnisse p > 0,2 unterstreicht. Die Farbe „orange" markiert Werte, die im Indifferenzbereich liegen: 0,05 < p < 0,2.

	Zahnarzt			Selbstbeurteilung		Differenz						
	N	M	SD	M	SD	M	SD	T	df	p (2-seitig)	r	p (2-seitig) von r
AEZ_Note	33	2,52	1,15	2,45	0,75	0,06	1,03	0,34	32	0,737	0,48	0,005
POA_Note	34	2,35	0,85	2,24	0,55	0,12	0,91	0,75	33	0,458	0,21	0,245
LAE_Note	34	2,85	0,89	2,59	0,66	0,26	1,11	1,39	33	0,173	0,00	0,986
ZE_Note	31	3,06	0,85	2,26	0,82	0,81	1,14	3,95	30	0,000	0,07	0,704
Naht_Note	33	2,91	0,63	2,91	0,95	0,00	0,71	0,00	32	1,000	0,67	0,000
BLS_Note	33	2,79	0,99	2,70	0,73	0,09	0,98	0,53	32	0,598	0,38	0,027
Gesamt-Note	34	2,74	0,50	2,53	0,45	0,21	0,52	2,32	33	0,027	0,39	0,021

Im Hinblick auf die Gesamt-Note ergibt sich zwischen den beiden Vergleichsgruppen eine Mittelwertedifferenz von 0,21 (SD 0,52). Der T-Wert liegt mit 2,32 hoch und der entsprechende p-Wert zeigt mit 0,027 signifikante Mittelwerteunterschiede. Zwischen Selbst- und Fremdeinschätzung besteht ein positiver Zusammenhang (r = 0,39), der signifikant ist, da p (r) = 0,021.

Vergleich der Bewertung der studentische Rater – Selbsteinschätzung der Studierenden

Zur Berechnung der Mittelwertedifferenzen werden die durch studentische Rater erhobenen Mittelwerte von den sich aus der Selbsteinschätzung ergebenen Mittelwerten subtrahiert. Für die Gesamtnote ergibt sich eine Mittelwertedifferenz von 0,01 (SD 0,53), die allerdings mit einem p-Wert von 0,932 nicht-signifikant ist, siehe Tabelle 15. Die Fremd- und die Selbsteinschätzungen korrelieren positiv, jedoch nicht hoch mit einem Korrelationskoeffizienten von r = 0,35. Die Korrelation ist mit p (r) = 0,045 signifikant. Es besteht folglich ein Zusammenhang zwischen der Selbsteinschätzung durch die Studierenden und der Fremdeinschätzung durch die studentischen Rater.

Tabelle 15: Vergleich studentischer Rater – Selbsteinschätzung

N = Stichprobengröße; M = Schwierigkeit/ Mittelwert; SD = Standardabweichung; T = T-Wert; df = Anzahl an Freiheitsgraden; p = p-Wert, r = Korrelationskoeffizient. Die Farbe „grün" weist auf ein signifikantes Ergebnis p < 0,05 hin, während die Farbe „rot" nicht signifikante Ergebnisse p > 0,2 unterstreicht. Die Farbe „orange" markiert Werte, die im Indifferenzbereich liegen: 0,05 < p < 0,2.

	studentischer Rater			Selbstbeurteilung		Differenz						
	N	M	SD	M	SD	M	SD	T	Df	p (2-seitig)	r	p (2-seitig) von r
AEZ_Note	33	2,33	0,82	2,45	0,75	-0,12	0,78	-0,89	32	0,379	0,51	0,003
POA_Note	34	2,59	0,78	2,24	0,55	0,35	0,77	2,66	33	0,012	0,37	0,031
LAE_Note	34	2,85	0,82	2,59	0,66	0,26	1,11	1,39	33	0,173	-0,12	0,515
ZE_Note	31	2,65	1,08	2,26	0,82	0,39	1,17	1,84	30	0,076	0,26	0,160
Naht_Note	33	2,52	0,87	2,91	0,95	-0,39	1,06	-2,14	32	0,040	0,32	0,066
BLS_Note	33	2,21	0,74	2,70	0,73	-0,48	0,83	-3,34	32	0,002	0,36	0,043
Gesamt-Note	34	2,54	0,48	2,53	0,45	0,01	0,53	0,09	33	0,932	0,35	0,045

Während die zahnärztlichen Rater insgesamt schlechter beurteilten, so bewerteten die studentischen Rater die Studierenden an 50% der Stationen schlechter und an 50% der Stationen besser als diese sich selbst einschätzten. Daraus ergibt sich eine sehr geringe Gesamt-Mittelwertedifferenz, die nicht signifikant ist. Die Gesamtnote zeigt eine positive Korrelation, die signifikant ist. Es besteht also ein Zusammenhang zwischen der studentischen Bewertung und der Selbsteinschätzung der Teilnehmenden.

3.2.4 Interkorrelation

Die Interkorrelation beschreibt den Zusammenhang zwischen dem Gesamtergebnis einer Station und dem Gesamtergebnis einer anderen Station für die jeweiligen Ratergruppen. Sie lässt Rückschlüsse auf die Qualität und Quantität des Zusammenhangs zwischen den Ergebnissen der beiden Stationen zu. Dieser Zusammenhang wurde für jede Station in Kombination mit jeder anderen Station erstellt, siehe Tabelle 16. Um den Zusammenhang darzustellen wird der jeweilige Korrelationskoeffizient ermittelt.

Tabelle 16: Interkorrelation zwischen den verschiedenen Stationen

*Oberhalb der Diagonalen sind die Korrelationen der zahnärztlichen Rater (Farbe „blau“) aufgelistet, unterhalb der Diagonalen die Korrelationen der studentischen Rater (Farbe „orange“). Die (mit * oder **) gekennzeichneten Werte ergeben die Signifikanzwerte. Ein Sternchen stellt dabei ein signifikantes Ergebnis dar (p < 0,05), zwei Sternchen ein sehr signifikantes Ergebnis (p < 0,01). Dies besagt, dass der zugehörige Wert nicht zufällig entstanden ist und zu 95% bzw. 99% nicht als Folge des Standardfehlers zu verstehen ist.*

	AEZ	POA	LAE	ZE	Naht	BLS
AEZ		0,181	-0,145	0,06	0,026	0,409*
POA	0,586**		0,061	0,137	0,146	0,128
LAE	0,067	0,155		0,381*	0,300	0,120
ZE	0,027	0,229	0,202		0,311	0,145
Naht	0,038	0,155	0,121	0,155		0,375*
BLS	0,025	0,123	-0,030	0,103	0,364*	

Ein negativer Korrelationskoeffizient beschreibt dabei eine negative lineare Funktion. Dies bedeutet, dass sich das Gesamtergebnis der einen Station verbessert, während sich das Gesamtergebnis der Vergleichsstation verschlechtert. Der Korrelationskoeffizient r = -1 beschreibt dabei einen perfekt negativ linearen Zusammenhang. Eine positive Korrelation gibt an, dass zwischen den beiden Stationen ein positiver, linearer Zusammenhang besteht und sich das eine Gesamtergebnis verbessert, wenn sich auch das andere verbessert. Ein Korrelationskoeffizient von r = 1 ergibt somit einen perfekt positiv linearen Zusammenhang. Nimmt r den Wert von 0 an, so ist kein Zusammenhang zwischen den beiden Gesamtergebnissen erkennbar.

Zwischen den beiden Gesprächsstationen unter der Bewertung der zahnärztlichen Rater besteht ein leicht positiver Zusammenhang. Dies wird anhand des Korrelationskoeffizienten von r = 0,181 ersichtlich. Die meisten Korrelationen zwischen den Stationen sind nicht-signifikant. Auffällig ist dabei die Kombination von Station 1 (AEZ) und Station 3 (LAE) mit einem Korrelationskoeffizienten von r = -0,145, der einen negativen Zusammenhang darstellt. Allerdings ist auch dieser Zusammenhang nicht signifikant. Als signifikant können drei der 15 Kombinationen beschrieben werden. Alle drei stellen einen positiven Zusammenhang dar und ergeben Signifikanzwerte von p < 0,05. Die erste signifikante Korrelation besteht zwischen den Stationen „LAE“ und „ZE“ mit r = 0,381. Ebenfalls signifikant sind die Korrelationen zwischen den Stationen „AEZ“ und „BLS“ (r = 0,409) und den Stationen „Naht“ und „BLS“ (r = 0,375).

Die Bewertung der studentischen Rater ergibt signifikante Kombinationen für die Stationen „AEZ“ und „POA“ und die Stationen „Naht“ und „BLS“. Die Korrelation der beiden Gesprächsstationen zeigt mit r = 0,586 den höchsten Korrelationskoeffizienten. Dieser ist signifikant ($p < 0,01$). Die Stationen „Naht“ und „BLS“ ergeben mit r = 0,364 ebenfalls einen positiven Zusammenhang, der signifikant ist.

3.2.5 Intra-Klassen-Korrelation

Die Bewertung der zahnärztlichen und der studentischen Rater ist hoch korreliert (ICC = 0,888). Die ICC der einzelnen Stationen schwanken zwischen 0,495 (mittelhohe Reliabilität) und 0,794 (hohe Reliabilität). Die zugehörige Tabelle ist im Anhang 7.4 zu finden. Die höchste ICC erhält die Station „Lokalanästhesie“ (0,794), die niedrigste die Station „postoperative Aufklärung“ (0,495). Für die Station „Zahnextraktion“ und die Station „Naht“ ergeben sich mit einer ICC von 0,737 und 0,705 ebenfalls hohe Reliabilitäten. Die Station „BLS“ und die Station „Aufklärung und Einwilligung Zahnextraktion“ haben eine ICC von 0,653 und 0,682. Die ermittelten Werte besagen, dass an Station „LAE“, „ZE“ und „Naht“ die höchsten Zusammenhänge zwischen den beiden Raterurteilen bestehen. Die Station „POA“ zeigt den geringsten Zusammenhang.

3.2.6 Beurteilung der tOSCE Software durch die Rater

Wie bereits im Material- und Methodenteil erwähnt, evaluierten beide Ratergruppen im Anschluss an die Hauptuntersuchung die Tablet-Nutzung. Sie beurteilten die Funktionalität der Software, die Softwareergonomie und die Fehlerrobustheit und gaben eine Gesamtbewertung ab, um daraus Verbesserungen für künftige Durchführungen ableiten zu können.

In Bezug auf die Funktionalität der Software lagen die Mittelwerte, die sich aus den relativen Häufigkeiten ergeben, im Bereich zwischen „trifft voll zu“ und „trifft überwiegend zu“. Auch die Software-Ergonomie zeigt Bewertungen in diesem Bereich. Drei der vier Fragen aus dem Bereich Fehlerrobustheit wurden gleichermaßen bewertet, nur das Kriterium „die Anwendung gibt mir ausreichend Hinweise, wie ich Fehler korrigieren kann“ zeigt Ergebnisse, die sich zwischen „trifft weitgehend zu“ und „trifft überwiegend zu“ befinden.

Die offenen Anmerkungen der Rater loben ebenfalls die Funktionalität des Scanners sowie der Software-Funktionen an sich und betonen deren einfache Handhabung.

Insgesamt überzeugt das Ergebnis der Software durch einfache Anwendung und hohe Funktionalität und Ergonomie. Es gab keine Probleme bei der Bewertung durch die Rater, und auftretende

Fehler konnten gut korrigiert werden. Für weitere OSCE-Prüfungen ist die Nutzung der Tablets demnach wünschenswert und trägt zur effizienten und einfachen Durchführung der Bewertung bei.

4. Diskussion

4.1 Die Modifikationen nach der Pilotuntersuchung

Aus den Ergebnissen der Pilot-OSCE konnten Änderungen zur Steigerung der Testgüte für die Hauptstudie vorgenommen werden: Zum einen wurden die Checklisten-Items mithilfe der Itemanalyse umformuliert und revidiert. Generell wurden die Items im Vergleich zur Pilot-OSCE in der Hauptstudie durch Positiv- und Negativbeispiele beschrieben, die für die Rater einsehbar waren. In der Pilot-OSCE war lediglich aufgeführt, wie die Studierenden das Item „richtig" umzusetzen hatten. An der Station „Aufklärung und Einwilligung Zahnextraktion" wurde das Item mit der niedrigsten Trennschärfe aus der Checkliste entfernt und die anderen durch detailliertere Formulierungen näher beschrieben. Die Auswirkung dieser Maßnahmen auf die Ergebnisse der Hauptstudie war positiv. Auch an der Station „Aufklärung post OP" wurden die Items durch detailliertere Beschreibungen geschärft. Der Checklisten-Mittelwert der Station und Cronbachs Alpha verbesserte sich daraufhin. An der Station „Lokalanästhesie" wurden ebenfalls einige Items präzisiert. Der Item-Mittelwert für das Item „nimmt die Leitung von Prämolaren der Gegenseite unter Knochenkontakt (es piept) vor" verbesserte sich durch die Elimination des Kriteriums „es piept" deutlich. Dies bezieht sich auf den Piepton des verwendeten Modells bei korrekter Injektion. Die Checkliste der Station „Zahnextraktion" hingegen wurde um ein weiteres Item ergänzt. Das Item „extrahiert den richtigen Zahn" wurde hinzugefügt, worauf die Station ein höheres Cronbachs Alpha zeigt. Die Korrelation zwischen Globalnote und Checklisten-Mittelwert zeigt für diese Station in der Pilot-OSCE einen Wert von -0,55, der nicht signifikant ist. Dies ist wie folgt zu erklären: Bei Extraktion des „falschen" Zahnes resultiert die Note „5". Da dies Kriterium jedoch kein Item der Checkliste ist, nimmt es auf diese auch keinen Einfluss. Das bedeutet, dass jemand, der den falschen Zahn extrahiert, zwar die Note „5" erhält, allerdings trotzdem gute Checklisten-Mittelwerte erhalten kann, da er die Kriterien angemessen erfüllt. Aufgrund dessen korrelierten Note und Skalenmittelwert an dieser Station nicht signifikant. Die Korrelation ist nach Abänderung in der Hauptstudie deutlich höher. An der Station „Naht" wurden die Items geschärft, zum Beispiel die Entfernung zum Wundrand und die Länge der Fadenenden mit Millimeterangaben präzisiert. Die Station zeigt in der Hauptstudie daraufhin verbesserte Werte für die Trennschärfen.

Die Zusammensetzung der Gesamtnote wurde nach der Pilot-OSCE modifiziert. Zuvor ergab sich die Gesamtnote aus einer Gewichtung von 80% Checklisten-Mittelwerten und 20% Globalnote. Auf Basis der Studienlage erweist sich ein globales Rating als gleichermaßen geeignet zur Evaluation der Testgüte einer OSCE wie das Checklisten-Rating (Nikendei und Jünger 2006). Für Prüfungen mit vorwiegend technischen Aufgabenstellungen wird das Checklistenformat als adäquat angesehen,

während kommunikative Fertigkeiten eher global bewertet werden (Newble 2004; Nikendei und Jünger 2006). Da in der Hauptstudie zwei der sechs Stationen kommunikative Fertigkeiten prüfen, setzt sich das Gesamtergebnis aus 70% Checklisten-Mittelwerten und 30% Globalnote zusammen. Hinzu kam außerdem eine sechste Station, die Station „Basic Life Support". Die Stichprobengröße erhöhte sich von N = 9 in der Pilotuntersuchung auf N = 36 in der Hauptuntersuchung. Die deskriptive Statistik ist mit einer Stichprobengröße von N = 9 nur bedingt aussagekräftig. Sobald mehrere Studierende das gleiche Ergebnis erzielen, hat dies bereits Auswirkungen auf die Verteilungseigenschaften, was bei größerer Stichprobenanzahl kaum Auswirkungen hat. Ebenso fällt ein Ausreißer bei einer kleinen Stichprobengröße mehr ins Gewicht als bei einer großen Stichprobe. Aufgrund der geringen Größe ist die Verteilung nicht resistent gegenüber Ausreißern. Damit verbunden sind auch Auswirkungen auf die Reliabilität. Aus diesem Grund besteht hier allein durch die Vergrößerung der Stichprobe die Möglichkeit, dass sich die Reliabilität verändert. Ein Einflussfaktor auf die Verteilungseigenschaften der Pilot-OSCE können auch die Studierenden selbst gewesen sein: „our sample is likely to contain a higher proportion of motivated students who were willing to spend extra time on study-related tasks" (Schiekirka et al. 2013, S. 374).

4.2 Beurteilung der Testgüte der Hauptuntersuchung

Die Validität gibt an, wie gut ein Test die Funktion erfüllt, das zu messen, was er vorgibt zu messen, die Reliabilität, wie genau der Test das misst, was er messen soll und die Objektivität, wie unabhängig der Test von den Personen ist, die ihn bewerten. Die jeweiligen Unteraspekte wurden bereits näher erläutert. Inwieweit entspricht nun die Testgüte den Anforderungen an eine summative OSCE-Prüfung?

Um diese Fragen zu klären, werden die einzelnen Items und Checklisten der jeweiligen Stationen näher betrachtet. Die zugehörigen Tabellen sind im Anhang unter 7.3.1 zu finden, der tatsächliche Inhalt der Items wurde aus Gründen der Wiederverwertbarkeit der Checklisten-Items codiert. Die Bestimmung der Testgüte ist wichtig zur Differenzierung und Evaluation „schlechter" Studierender. Einige Items zeigen bei einer niedrigen Aufgabenschwierigkeit hohe Trennschärfen. Diese Ergebnisse sagen aus, dass obwohl eine große Anzahl an Kandidaten die Aufgabe „richtig" erfüllte, trotzdem gut zwischen leistungsstarken und leistungsschwachen Studierenden unterschieden werden konnte. Die Kandidaten, die dieses Item zufriedenstellend lösten, schnitten demzufolge auch ansonsten gut ab. Ein Schwierigkeits-Wert von 0,931 sagt aus, dass 93,1% der Kandidaten zu dieser Gruppe Studierender gehören. Die Aussagekraft dieses Ergebnisses ist nur bedingt nützlich. Mittlere Schwierigkeitswerte sind daher sehr viel aussagekräftiger, da hier die Trennschärfe maximale Werte

erreichen kann. Anhand der Trennschärfe kann bestimmt werden, ob eine Aufgabenstellung verändert werden sollte oder beibehalten werden kann, oder aber eliminiert werden sollte (Schelten 1980). Hohe Item-Mittelwerte, also leichte Aufgabenstellungen, sind zwar nur bedingt aussagekräftig in Bezug auf die individuellen Leistungen der Studierenden, allerdings werden erst durch die Trennschärfe die Konsequenzen für das jeweilige Item klassifiziert. Trennschärfen unter 0,19 gelten als „unbrauchbar“ und erfordern eine Schärfung. Liegt die Trennschärfe zwischen 0,20 und 0,29, so sollte das Item verbessert werden. Items, die eine Trennschärfe von 0,30 bis 0,39 zeigen sind zwar brauchbar, es sollte allerdings über Verbesserungsmöglichkeiten nachgedacht werden, um dann eine Trennschärfe von 0,40 oder höher zu erreichen, die als ausgezeichnet gilt (Schelten 1980, S. 135).

Items mit einer geringen Schwierigkeit und einer brauchbaren Trennschärfe sollten nicht unbedingt eliminiert werden, obwohl ihre Aussagekraft, gemessen an der Schwierigkeit, niedrig ist. Sie können auch als Items angesehen werden, die von Studierenden in der Regel ohne Probleme beantwortet werden können und sich somit positiv auf die Auswertung auswirken. Außerdem kann es gewollt sein, dass einige Items von allen Studierenden richtig erfüllt werden: „Im Sinne einer kriteriumsorientierten Prüfung ist es jedoch durchaus sinnvoll, wichtige basale Fertigkeiten oder Kenntnisse in einer Prüfung unabhängig von ihrer erwarteten Schwierigkeit abzufragen“ (Möltner et al. 2006, S. 3). Durch die Umgestaltung von Items mithilfe zusätzlicher Kriterien können diese „schwerer gemacht“ werden, um die empfohlene Aufgabenschwierigkeit von Möltner zu erfüllen (Möltner et al. 2006). Dies gilt beispielsweise für Items, die einen Mittelwert von 1,0 und zugehöriger Trennschärfe von 0,0 ergeben.

Die Itemanalyse für die Bewertung durch die studentischen Rater wird an dieser Stelle vernachlässigt, da die Bewertung durch die ärztlichen Rater in dieser OSCE ausschlaggebend für die Bewertung der Prüflinge ist.

4.2.1 Die Testgüte der beiden kommunikativen Stationen

Die Station „Aufklärung und Einwilligung Zahnextraktion“ zeigt eine überzeugende interne Konsistenz der Checkliste, sowie des Gesamtergebnisses. Die Station kann bereits als reliabel bezeichnet werden.

Es gibt für die Bewertung durch die ärztlichen Rater zwei auffällige Items. Sie fallen durch geringe Item-Mittelwerte und hohen Trennschärfen auf, siehe Tabelle 23. Der Inhalt dieser beiden Items steht chronologisch gesehen ganz am Ende des Patientengesprächs, so zum Beispiel der Erhalt der schriftlichen Einwilligung nach der Aufklärung. Aus diesem Grund kann angenommen werden, dass das Ergebnis nicht aufgrund von inhaltlichen Defiziten der Studierenden zustande kommt. Vielmehr müssen hier andere Ansätze in Betracht gezogen werden: Die Station ist also entweder zu lang

konzipiert, die Zeit zur Bearbeitung der Aufgabenstellung zu kurz gewählt, die Studierenden zu langsam in ihrer Ausführung oder die SP mangelhaft instruiert. Die hohen Trennschärfen tragen zur guten Reliabilität der beiden Items bei und unterstreichen dieses Argument. Studien zeigen, dass bei der Auswertung von Aufgabenstellungen Vorsicht geboten ist, bei denen die Prüflinge „auf Grund der beschränkten Prüfungszeit nicht alle Aufgaben bearbeiten konnten“ (Möltner et al. 2006). Das ist aber nicht damit gleichzusetzen, dass der Inhalt nicht adäquat geprüft wurde. "Auf Grund von Zeitmangel nicht bearbeitete Aufgaben sind bei den Prüflingen von der Bewertung auszuschließen" (Möltner et al. 2006, S. 3).

An der Station „postoperative Aufklärung“ sind über nahezu alle Items hohe ausgezeichnete Trennschärfen vorhanden. Die Konsistenz der Checkliste kann als sehr reliabel bezeichnet werden und die Reliabilität zeigt sowohl für die Checkliste als auch für das Gesamtergebnis aus Checkliste und Globalnote einen hohen Alpha-Koeffizienten. An der Station fallen ebenfalls zwei Items durch niedrige Trennschärfen auf, siehe Tabelle 24. Die beiden Items prüfen Aspekte der Zahnarzt-Patienten-Beziehung ab und somit die soziale Kompetenz - wie Empathie - der Studierenden als essentiellen Bestandteil ihrer medizinischen Rolle (Frank 2005). Sie sind also inhaltlich nicht rein objektiv zu bewerten und dennoch von besonderer Bedeutung für ihr Rollenverständnis. Es ist zudem schwierig, zwischen tatsächlicher Empathie und zu Prüfungszwecken „vorgespielter“ Empathie zu differenzieren. Dies kann von den unterschiedlichen Ratern unterschiedlich empfunden werden. Das Item zeigt folglich keine stabile Auswertungsobjektivität für die OSCE-Station. Zur dauerhaften Etablierung in die Checkliste dieser OSCE-Station sollten die psychosozialen Aspekte aus der Item-Beschreibung entfernt werden. Die zum Gesamteindruck zählenden Inhalte dieses Items werden bereits durch die Vergabe der Globalnote erfasst. In diese gehen auch maßgeblich Verhaltensweisen mit ein, wie z.B. Empathie, die durch die Checklisten-Items unberücksichtigt bleiben (Chenot et al. 2007). Somit bleiben diese Aspekte weiterhin berücksichtigt.

Item 11 zeigt ebenfalls eine niedrige Trennschärfe und prüft inhaltlich die Handhabung von Komplikationen bzw. des normalen Heilungsverlaufs der Wunde. Ergänzt werden sollte an dieser Stelle, nach wie vielen Tagen genau in beiden Fällen zu handeln ist. Damit wird sicher ausgeschlossen, dass sich die Trennschärfen nicht aufgrund unzureichend detaillierter Beschreibung des Items und unterschiedlicher Auslegungen dieser Formulierung durch die Rater ergeben.

4.2.2 Die Testgüte der drei praktisch zahnärztlich-chirurgischen Stationen

Station Lokalanästhesie

Diese Station zeigt überwiegend hohe Item-Mittelwerte, also relativ „leichte“ Aufgabenstellungen. Die Station zeigt eine unbrauchbare Checklisten-Trennschärfe und die Mehrheit der Items verbesserungsbedürftige Trennschärfen. Zwei davon sind mit Trennschärfen im negativen Bereich kritisch zu betrachten. Beide Items beschreiben grundlegende Kenntnisse über die Gabe einer Lokalanästhesie und können somit nicht aus der Checkliste entfernt werden. Item 11 ist klar und präzise formuliert und beschreibt die Art und Weise der Einstichtechnik beim Legen einer Lokalanästhesie. Es beinhaltet jedoch, dass Knochenkontakt vorgenommen wird. An dieser Stelle ist das Item möglicherweise nicht ausreichend valide gestaltet. Da die Studierenden während der Prüfungszeit die Anweisungen haben, nicht zu sprechen, ist es dem Rater nicht möglich herauszufinden, ob der Studierende wirklich Knochenkontakt hat, ohne ihn zu befragen. Eine Befragung durch die Rater ist jedoch nicht erwünscht. Sofern gesichert ist, dass der Rater ausreichend Sicht auf das Gebiet hat ist es möglich, zu differenzieren aus welcher Richtung der Studierende einsticht und wie weit er die Kanüle vorschiebt, ob er allerdings Knochenkontakt hat, ist nicht mit ausreichender Sicherheit zu beurteilen. Das Item gestaltet sich aufgrund dessen nicht ausreichend valide. Andere Items beinhalten Formulierungen wie „hygienisch einwandfreies Verhalten“ (Item 4), „lagert…sinnvoll“ (Item 5) oder „aspiriert langsam“ (Item 13), siehe Tabelle 25. Alle diese Formulierungen lassen Raum für unterschiedliche Interpretationen der Rater und zeigen keine gute Auswertungsobjektivität. Die Items müssen für die nächste OSCE umformuliert und objektiver gestaltet werden. Da die Analyse für viele der Items unbrauchbare Trennschärfen ergibt, muss an dieser Station über eine alternative Durchführung nachgedacht werden. Um zu ermitteln, ob die Ergebnisse aufgrund eines externen Faktors zustande kommen oder Resultat der mangelnden theoretischen Grundkenntnisse der Studierenden zum Thema Lokalanästhesie sind, wäre es z.B. sinnvoll, die Aufgabenstellung zu verändern und zu integrieren, dass die Studierenden während der praktischen Übung zusätzlich beschreiben dürfen, was sie tun. Auf diese Art und Weise ist es den Ratern möglich, zu bewerten, inwieweit der Durchführung des Studierenden praktische Kenntnisse zugrunde liegen. Die Reliabilität kann derzeit nur als mittelmäßig beschrieben werden. Zusammen mit dem Gesamteindruck gilt sie aber nach Tudiver bereits als akzeptabel (Tudiver et al. 2009).

Station Zahnextraktion

Diese Station ergibt überwiegend hohe Item-Mittelwerte, allerdings auch viele gering trennscharfe Items. Dennoch zeigen die Werte für die Checklisten-Trennschärfe und Cronbachs Alpha keine zufriedenstellenden Ergebnisse, siehe Tabelle 26.

Einige Items sind noch nicht ausreichend auswertungsobjektiv gestaltet. Zum Beispiel kann Item 4 („legt alles Benötigte sinnvoll bereit“) von unterschiedlichen Betrachtern unterschiedlich ausgelegt werden. Es sollte zum einen definiert werden, was genau zu den benötigten Dingen gehört, die der Studierende bereitlegen soll, zum anderen inwieweit es „sinnvoll“ bereitzulegen ist. Zwei der Items ergeben kritische Trennschärfen. Bei beiden Items ist der Ratereinfluss ein Aspekt der diskutiert werden sollte: Item 7 erfasst inhaltlich den Schutz des Hebels während der Arbeit und ergibt ein stark negativ trennscharfes Ergebnis. Auch Item 11 beinhaltet Grundsätze zur sicheren Durchführung der Extraktion, diesmal allerdings zur Stabilisierung des Alveolarfortsatzes während der Extraktion. Der Inhalt beider Items ist im Hinblick auf die Auswertungsobjektivität nicht präzise genug gestaltet und sollte geschärft werden.

An dieser Station stellen die Sichtweisen der Rater und somit die Ratereinflüsse ein wichtiges Kriterium dar. Diese können Ansatzpunkt zur Steigerung der Testgüte sein, ebenso wie eine Schärfung der Items. Die Korrelation zwischen Checklisten-Mittelwert und Globalnote dieser Station steigt nach der Überarbeitung durch die Ergebnisse der Pilot-OSCE deutlich an. Dies ist Resultat der Aufnahme des Kriteriums „extrahiert den richtigen Zahn“.

Station Naht

Die Nahtstation ergibt ebenfalls überwiegend hohe Item-Mittelwerte. Es liegen keine negativ trennscharfen Items vor, siehe Tabelle 27. Auch in dieser Checkliste sind Begriffe zu finden, die im Hinblick auf die Auswertungsobjektivität kritisch zu betrachten sind. Sie enthalten Wörter wie „korrekt“ oder „richtig“, die nicht näher beschrieben werden. Dennoch sollen die Rater nur anhand dieser Beschreibung das Kriterium bewerten. Es muss hier sichergestellt werden, dass diese Items entweder objektiviert werden, indem ausgeführt wird, was sie genau beinhalten, oder die Rater im Rahmen der Raterschulung eine genaue Anweisung dazu erhalten. Was genau ist „richtig“ und was ist „falsch“? Die erste Option ist an dieser Stelle als geeigneter zu bewerten, da sie weniger fehleranfällig ist. Es ist möglich, dass Aspekte, die in der Raterschulung besprochen wurden zu Beginn der OSCE-Prüfung von den Ratern nicht mehr in vollem Umfang abgerufen werden können. Es ist sinnvoll die Präzisierungen schriftlich in die Checkliste zu integrieren, sodass es den Ratern während der Prüfung möglich ist, diese erneut abzurufen. Die sicherste Methode ist sicherlich eine Kombination aus beiden Merkmalen. Die Rater müssen bereits während der Raterschulung ausführlich über die einzelnen

Items der Stationen informiert werden und ihnen muss die Möglichkeit eingeräumt werden auftretende Fragen sofort vor Ort zu klären. Gleichzeitig unterstützt eine Präzisierung der Items in den Checklisten auf dem Tablet die Rater während der Prüfung.

Einige der Items zeigen sehr hohe Item-Mittelwerte und dazugehörige geringe Trennschärfen. Diese Items sollten folglich differenziert betrachtet und auf ihre Inhaltsvalidität geprüft werden. Item 11 liegt im negativ trennscharfen Bereich und legt eine Eliminierung des Items nahe. Dennoch enthält das Item inhaltlich sehr wichtige grundlegende Aspekte zum hygienischen Verhalten des Studierenden während der „Behandlung". Es darf daher aus zahnmedizinischer Sicht nicht eliminiert werden. In Kombination mit dem Gesamteindruck zeigt die Station Tendenzen in Richtung „guter" interner Konsistenz.

4.2.3 Die Testgüte der Station Basic Life Support

Diese Station zeigt eine sehr geringe Reliabilität und schlechte statistische Ergebnisse, weshalb der Sinn dieser Station diskutiert werden muss. Aufgrund der generalisierten niedrigen statistischen Ergebnisse, die mit einer geringen Reliabilität der Checkliste und der gesamten Station einhergehen, ist eine Eliminierung der Station in Betracht zu ziehen. Es ist hier offensichtlich, dass die Testgütekriterien nicht erfüllt sind. Die Station ist zwar aufgrund der standardisierten Checkliste objektiv gestaltet, sie liefert allerdings im Hinblick auf Reliabilität und Validität keine akzeptablen Ergebnisse. Die Ergebnisse der Station lassen mehrere Erklärungsansätze zu. Die Formulierung der zwölf Checklistenitems ist größtenteils sehr präzise und auswertungsobjektiv gestaltet und lässt den Beurteilern keinen Spielraum für Interpretationen. Zwei Items zeigen dahingehend Auffälligkeiten: Das Kriterium ist mit dem Wort „korrekt" beschrieben und dieser Begriff wurde nicht weiter präzisiert. „Korrekt" ist ebenfalls nicht ausreichend auswertungsobjektiv, da es subjektiv ausgelegt werden kann, z.B. „korrekte Entlastung". Dieses Item ergibt in der Itemanalyse eine negative Trennschärfe von -0,256, siehe Tabelle 28. Das Kriterium ist unbrauchbar für die Checkliste, da es nicht dazu beiträgt zwischen den leistungsstarken und den leistungsschwachen Studierenden zu differenzieren. Vielmehr besagt dieser Wert, dass das Item etwas anderes misst als der Rest der Station. Die Qualität dieses Items ist folglich sehr schlecht (Kiehl et al. 2014), und es muss zwingend umgestaltet oder gestrichen werden. Eine korrekte Entlastung an dieser Station bedeutet, dass der Brustkorb des Simulationspatienten zwischen zwei Pumpstößen der Herzdruckmassage komplett entlastet wird, damit sich das Herz in der Zeit mit Blut füllen kann, ohne dass der Druckpunkt dabei aufgegeben wird. Das Item könnte also heißen „komplette Entlastung des Brustkorbs zwischen jeder Thoraxkompressionen unter Beibehaltung des Druckpunkts". Entsprechendes gilt für das Item „Druckpunkt

korrekt“. Der korrekte Druckpunkt befindet sich in der Mitte des Brustkorbs und dies sollte zur Verdeutlichung auch in das Item aufgenommen werden, da auch dieses Item kein trennscharfes Ergebnis liefert. Um die Station reliabel zu gestalten muss sichergestellt werden, dass die Rater ausreichend Sicht auf den Phantompatienten haben. Dieser hat an seiner unteren Seite eine Anzeige, die dem Betrachter die Entscheidung zur korrekten Drucktiefe und Frequenz erleichtert. Es ist fraglich, ob alle Rater während der OSCE-Prüfung auch gute Sicht auf diese Anzeige hatten.

Die BLS Station liefert in diesem Zustand keine Ergebnisse, welche es zulassen diese Station in einer summativen Prüfung einzusetzen. Zudem wurde seit dem Sommersemester 2013 der „Basic Life Support“ in den Operationskurs integriert. Die BLS Station soll zu diesem Zweck als Sicherung der Leistungsverbesserung durch diese curriculare Veränderung dienen. Außerdem ist es auch für den Alltag der Zahnmediziner wichtig diese lebenserhaltenden Skills routiniert ausüben zu können, um eventuell auftretende Notfälle in der Praxis sicher handhaben zu können. Aus den genannten Gründen soll diese Station weiterhin Teil der OSCE Prüfung bleiben. Es muss daher sichergestellt werden, dass diese Station so weit verbessert werden kann, dass sie in einer summativen OSCE-Prüfung Bestand zeigt. Um die Testgütekriterien der Station zu verbessern, müssen die genannten Items der Checkliste umformuliert und präzisiert werden. Es sollte sichergestellt werden, dass die Rater ausreichend Sicht auf die Prüfungssituation haben, um diese bestmöglich bewerten zu können. Die Studierenden können mit zusätzlichem Vorbereitungsmaterial für diese Station sensibilisiert werden.

4.2.4 Zusammenfassung

Nach Überarbeitung der dargestellten Inhalte der Checklisten sind die Stationen ohne weiteres übertragbar auf eine als summatives Prüfungsformat durchzuführende OSCE.

Eine Schärfung der Checklisten kann im Allgemeinen auf verschiedene Art und Weise erreicht werden. Die einzelnen Items können revidiert werden, also in ihrer Formulierung verändert, gekürzt oder um detailliertere Beschreibungen erweitert. Wie genau diese Veränderungen aussehen sollen ist für jedes Item differenziert zu betrachten. Worte, die den Ratern Spielraum für subjektive Auslegungen lassen und die Aufgabenstellung der Studierenden beschreiben, wie „sinnvoll“, „gut“ oder „nachvollziehbar“ sollten zur Schärfung der Checkliste objektiviert werden. Die Hauptstudie zeigt, dass es außerdem sinnvoll ist, die Items anhand von Positiv- und Negativbeispielen zu beschreiben, also den Ratern schriftlich aufzuführen, was als eine „gute“ Leistung zu bewerten ist und was als eine „schlechte“ Leistung. Dies trägt zur Klarheit der Aufgabenstellung bei. Es ist bekannt, dass die Checkliste auf dem Fundament "eine[r] sorgfältige[n] Aufgabenauswahl

(Vermeidung von Konstruktunterrepräsentation durch eine hinreichend breite und repräsentative inhaltliche Abdeckung des Themengebiets und Vermeidung von konstruktirrelevanter Varianz durch Klarheit der Aufgabenstellung)" (Möltner et al. 2006, S. 11) erstellt werden soll. Dazu zählt ebenfalls die Entfernung von Checklisten-Items, die menschliche Kompetenzebenen - wie Empathie - testen (Kassam et al. 2016). Die Bewertung dieser Skills erfolgt dafür mithilfe der Vergabe einer Globalnote (Nikendei und Jünger 2006). Die Ratereinflüsse sollten ebenfalls in Betracht gezogen werden und die Raterschulung unter deren Berücksichtigung modifiziert und intensiviert werden.

4.3 Diskussion der Hypothesen

4.3.1 Wiederholung der Hypothesen

Die in der Einleitung dargestellten Hypothesen sollen an dieser Stelle kurz wiederholt werden:

(1) Die Nullhypothese H_1 geht davon aus, dass es keine signifikanten Mittelwerteunterschiede zwischen den Urteilen der zahnärztlichen Ratern (A) und der studentischen Ratern (S) gibt. Die Alternativhypothese H_{A1} besagt, dass signifikante Mittelwerteunterschiede zwischen den beiden Fremdurteilen vorhanden sind: H_1: ZR = SR; H_{A1}: ZR ≠ SR.

(2) Die Nullhypothese H_2 geht davon aus, dass sich die Raterurteile einer Ratergruppe untereinander nicht signifikant unterscheiden. Die Alternativhypothese H_{A2} besagt hingegen, dass signifikante Unterschiede zwischen der Bewertung der zahnärztlichen und der studentischen Rater bestehen: H_2: R1 = R2 und H_{A2}: R1 ≠ R2.

Die bereits erläuterten Kriterien lassen Aussagen über die Eignung der Checklisten der einzelnen Stationen zu. Die ideale Station vereint die Anforderungen an hohe Validität, Reliabilität und Objektivität und bietet dem Studierenden somit objektive und standardisierte Prüfungsbedingungen klinisch-praktischer Fertigkeiten. Um die Stationen in Zukunft als Teil einer formativen, oder auch summativen OSCE-Prüfung einsetzen zu können, müssen also ebenfalls die Stärken und die Schwächen der Methodik erörtert werden.

4.3.2 Stärken-Schwächen-Analyse der Methodik

4.3.2.1 Die Checklisten zur Bewertung der OSCE

Die einzelnen Items wurden inhaltlich valide gestaltet und prüfen das Wissen ab, das den Studierenden in den zahnärztlich-chirurgischen Operationskursen der ersten drei klinischen Semester vermittelt wurde. Die Erstellung eines Blueprints gewährleistet, dass die Curriculumsziele angemessen repräsentiert werden und trägt maßgeblich zur Sicherung der Inhaltsvalidität bei (Nikendei und Jünger 2006; Davis 2003). Es wurde bereits deutlich, dass die äußeren Rahmbedingungen der Prüfung Einfluss auf die Validität nahmen. Eine mangelnde Umsetzung der Inhaltsvalidität wurde aufgrund von undeutlichen Formulierungen der Items erkennbar, wodurch die Checkliste an Inhaltsvalidität verliert. Dieser Punkt ist als Schwäche der Methodik zu betrachten. Die Aussage darüber, ob der Test auch das misst, was den Studierenden gelehrt wurde ist nur dann zuverlässig zu treffen, wenn die äußeren Rahmbedingungen für die Prüfung stimmen. Die Inhaltsvalidität ist allerdings meistens subjektiv geprägt. Aus diesem Grund ist es auch wichtig, zu überprüfen, inwieweit die Konstruktvalidität durch die Checkliste repräsentiert wird.

Die Konstruktvalidität gibt an, inwieweit der Test die Lehrinhalte, die dieser Station zugrunde liegen, widerspiegelt. Über die Interkorrelationen können Aussagen über die Zusammenhänge eines Konstruktes mit einem anderen Konstrukt, also einer OSCE-Station mit einer anderen OSCE-Station, getroffen und empirisch ausgedrückt werden. Anhand der Daten, die im Ergebnisteil beschrieben wurden, lässt sich empirisch zeigen, dass den Stationen unterschiedliche Konstrukte zugrunde liegen, siehe Ergebnisse 3.2.4. Die Interkorrelationen sind größtenteils gering und nicht-signifikant, die Stationen stehen also überwiegend nicht in Zusammenhang. Die Konstruktvalidität ist folglich hoch. Die Stationen messen verschiedene Aspekte. Dies geht mit einer hohen Variabilität der Aufgabenstellungen einher und ist eine Stärke dieser OSCE-Prüfung. Im Rahmen der OSCE-Prüfung ist es sinnvoll, möglichst viele unterschiedliche Facetten zu prüfen und den Aufgabenstellungen möglichst unterschiedliche Konstrukte zugrunde zu legen, um die Prüfung vielseitig zu gestalten. Es ist nicht wünschenswert, dass eine OSCE-Prüfung nur aus Stationen besteht, die dasselbe Konstrukt erfassen. Eine OSCE-Prüfung kann sogar erst dann als hochwertig bezeichnet werden, wenn sie mehrere Konstrukte abprüft (Möltner et al.2006). Dies kann für die formative OSCE-Prüfung der Mund-, Kiefer und Gesichtschirurgie bestätigt werden.

Die Reliabilität der OSCE und die interne Konsistenz der Checklisten kann als „gut“ bewertet werden. Maßnahmen zur Steigerung der Trennschärfe liegen in der Revision, Schärfung und Umformulierung einzelner Items. Das Ergebnis kann dadurch noch verbessert werden. Auch durch Variationen der „Testlänge“ kann Einfluss auf die Reliabilität der OSCE genommen werden. Dies bezieht sich auf

die Anzahl der Stationen an sich, nicht aber auf die Anzahl an Items einer Checkliste, da "eine steigende Zahl von Items die Reliabilität und Validität des OSCE reduziert" (Nikendei und Jünger 2006, S. 2). In Anbetracht der zur Verfügung stehenden Zeit ist es für den Rater mit steigender Anzahl an Items schwieriger, die Übersicht zu behalten, besonders da diese in den seltensten Fällen chronologisch von den Studierenden abgearbeitet werden. So kann eine Checkliste mit zu vielen Items schnell unübersichtlich werden. Sofern ein Item nicht ausreichend inhaltsvalide ist und unbrauchbare Trennschärfen zeigt, sollte dieses Item deshalb eher eliminiert als umformuliert oder revidiert werden. Je länger der Test konzipiert ist, also je mehr Stationen eine OSCE-Prüfung beinhaltet, desto höher wird die entsprechende Reliabilität. Aus der Literatur ist zu entnehmen, dass zur Durchführung einer hochwertigen, summativen OSCE-Prüfung 10-14 Stationen benötigt werden (Nikendei und Jünger 2006; Chenot und Ehrhardt 2005). Die Steigerung der Stationszahl hat also ebenfalls positive Auswirkungen auf die Höhe der Reliabilität. Um eine hochwertige OSCE-Prüfung in der UMG etablieren zu können müssen ohnehin mindestens weitere sechs Stationen konzipiert werden. Die Anzahl der Stationen dieser OSCE ist folglich noch eine Schwäche und hat einen negativen Effekt auf die Reliabilität, die mit der Etablierung weiterer Stationen, so wie es auch zukünftig geplant ist, erhöht werden kann.

Der dritte Aspekt, der zur Verbesserung der Reliabilität der Stationen beitragen kann, ist die Homogenität der Aufgabenstellungen. Je inhaltshomogener die Aufgaben einer Checkliste sind, desto größer wird die entsprechende Reliabilität. An den beiden Kommunikations-Stationen werden zum einen die inhaltlichen theoretischen Kenntnisse der Studierenden geprüft, als auch psychosoziale Aspekte wie Patientenumgang. Diese beiden Aspekte zeigen zwei unterschiedliche inhaltliche Schwerpunkte, die jedoch unmittelbar miteinander in Zusammenhang stehen. Homogener ließe sich die Station gestalten, indem Aspekte wie „Empathie“ und „Zuwendung“ durch die Vergabe der Globalnote abgedeckt und aus der Checkliste entfernt werden. Die anderen Stationen zeigen bereits inhaltlich homogene Checklisten. Auch die Aufgabenstellungen, die für die Prüflinge während der OSCE-Prüfung sichtbar sind, gestalten sich entsprechend homogen. Aufgrund dessen sollte dies vorerst so beibehalten werden.

Den Studierenden standen die gleichen Aufgabenstellungen, die gleichen Hilfsmittel und die gleichen Rahmenbedingungen zur Verfügung. Sie absolvierten die Aufgaben in derselben Zeitspanne, erhielten keinerlei Hilfestellung von externen Personen und durchliefen die Stationen in derselben Reihenfolge. Dies deutet auf eine hohe Durchführungsobjektivität und eine Stärke der Methodik hin.

Die Auswertungsobjektivität wird mithilfe der Standardisierung der Checklisten umgesetzt. In Kombination mit der Raterschulung schaffen die ausführlich formulierten Checklisten standardisierte Bedingungen für sowohl zahnärztliche und ärztliche als auch studentische Rater. Die erhobenen Ratereffekte bestätigen, dass keine signifikanten Ratereinflüsse vorhanden sind und die

Rater unabhängig voneinander zu ähnlichen Ergebnissen kommen. Dies gilt für ärztliche und studentische Rater gleichermaßen.

Da es sich um einen standardisierten Test handelt, kann die Interpretationsobjektivität vernachlässigt werden. Der Maßstab, auf dem der Vergleich beruht, ist hier gleich. Das bedeutet, dass zwangsläufig das gleiche Ergebnis resultiert. Die Werte, die sich für die eine Station ergeben, können direkt mit Werten einer anderen verglichen werden, beispielsweise Trennschärfen, Cronbachs Alpha oder dem mittleren Schwierigkeitswert.

4.3.2.2 Die Effizienz der Raterschulung

Die Ergebnisse der Bravais-Pearson-Korrelation und der Intra-Klassen-Korrelation zeigen hohe und signifikante Zusammenhänge zwischen den Bewertungen der studentischen und den Bewertung der ärztlichen Rater. Auch im Rahmen der Mittelwertevergleiche wird deutlich, dass überwiegend nichtsignifikante und allgemein geringe Mittelwerteunterschiede sowohl für die Checklistenbewertung als auch für die Globalbewertung und das Gesamtergebnis zwischen den beiden Fremdurteilen festzustellen sind.

Die Ergebnisse entstehen auf Basis der im Vorfeld stattfindenden Raterschulung, an der sowohl zahnärztliche und ärztliche als auch studentische Rater gleichermaßen teilnahmen. Die Aufgabe der Raterschulung lautet primär gleiche Grundvoraussetzungen für die beiden Ratergruppen zu schaffen. Sie wurden instruiert keine Hilfestellungen zu geben und auf die wichtigen Aspekte der jeweiligen Stationen vorbereitet. Auch mit den Räumlichkeiten, in denen die OSCE-Prüfung stattfinden sollte, wurden die Rater zuvor vertraut gemacht. Die Einsicht in die Videoaufnahmen der Prüfungsinhalte, sowie in die Kriterien der Checklisten tragen zur Standardisierungen der Prüfungsbedingungen bei.

Die studierenden Rater bewerteten die Studierenden weder signifikant besser noch signifikant schlechter als die ärztlichen Rater. Dies konnte bereits in anderen Göttinger Studien gezeigt werden (Chenot et al. 2007). Es ergaben sich weder signifikante Unterschiede in den jeweiligen Bewertungen, noch signifikante Ratereinflüsse zwischen den ärztlichen Ratern untereinander sowie den studentischen Ratern unter sich. Zudem waren hohe und signifikante Zusammenhänge zwischen den beiden Raterurteilen zu erkennen. Dies bestätigt, dass mithilfe einer guten Schulung eine hohe Interraterreliabilität zwischen den ärztlichen und den studentischen Ratern in der Zahnmedizin erreicht werden kann, die nicht vom Stand der Ausbildung abhängig ist (Wilkinson et al. 2003). Die Schulung ist eine Stärke der Prüfung, die im Rahmen der Vorbereitungen auf die Prüfung durchgeführt wurde. Sie ist Grundlage der bereits dargestellten statistischen Ergebnisse dieser OSCE.

Studien bestätigen, dass es sich bewährt, mit den Ratern vorab eine intensive Schulung durchzuführen (Boulet et al. 2003; Chenot et al. 2007; Friedman 2000). Teilweise werden noch intensivere Vorbereitungsmethoden zur Schulung der Rater empfohlen als sie in dieser OSCE vorgenommen wurden. Ziel dabei ist es, ein möglichst genaues Bild des Prüfungsablaufs zu vermitteln. Dazu gehört auch die Durchführung eines Beispieldurchlaufs (Boulet et al. 2003). Im Rahmen dieses Durchlaufs wird eine Prüfungssituation nachgestellt und die Rater können sowohl die Handhabung mit den Tablets testen als auch eine Vorstellung der Inhalte der Stationen und der Checklisten-Items bekommen. Diese Art der intensiveren Raterschulung birgt jedoch zwei Probleme: Zum einen gestaltet sich eine derartige Raterschulung zeitintensiver und belastet die personellen Ressourcen. Zum anderen müssen freiwillige Studierende gefunden werden, die sich für diese Übungen zur Verfügung stellen und mit deren Hilfe die Rater die OSCE-Prüfung simulieren können. Beide Aspekte ziehen höhere logistische und organisatorische Aufwände mit sich. Es wäre dennoch sinnvoll und zielführend die Schulung langfristig in diese Richtung zu entwickeln. Alternativ kann die Qualität der Schulung auch durch Video-basiertes Schulungsmaterial verbessert werden. An den praktischen zahnärztlich-chirurgischen Stationen wurde bereits mit derartigem Material gearbeitet, die Inhalte basierten jedoch auf Demonstrationen der Lehrkörper. Zusätzlich dazu wäre es sinnvoll den Ratern ein positives und ein negatives Beispielvideo einer vergangenen OSCE-Prüfung zu Verfügung zu stellen, um sie zu schulen (Friedman 2000).

4.3.3 Zusammenfassende Beurteilung der Hypothesen

Die beiden Ratergruppen unterscheiden sich kaum in ihren Mittelwerten, zeigen jedoch hohe signifikante Zusammenhänge zwischen den beiden Fremdurteilen. Die Korrelation der Gesamt-OSCE ist signifikant und mit $r = 0,9$ am höchsten. Dies bestätigt die Nullhypothese H_1, die Alternativhypothese H_{A1} wird somit verworfen.

Auch für den Vergleich der zahnärztlichen/ärztlichen und der studentischen Rater untereinander zeigen sich keine signifikanten Ergebnisse. Auch hier wird die Nullhypothese H_2 bestätigt. Es gab sowohl für die zahnärztlichen/ ärztlichen Rater, als auch für die studentischen Rater keine signifikanten Ratereinflüsse (Chenot et al. 2007; Ogden et al. 2000). Die Alternativhypothese H_{A2} wird somit verworfen.

Trotz der statistischen Ergebnisse ist es an allen Stationen möglich, weitere Verbesserungen zur Steigerung der Testgüte vorzunehmen. Die Stationen zeigen sowohl eine hohe Augenscheinvalidität als auch eine hohe Konstruktvalidität aufgrund der unterschiedlichen Konstrukte, die den Stationen zugrunde liegen. Diese konnten anhand der Interkorrelationen empirisch festgestellt werden. Die

OSCE-Prüfung misst unterschiedliche Facette der einzelnen Stationen und erhöht somit ihre Wertigkeit. Dies ist eine Stärke der angewandten Methodik. Die Reliabilität der OSCE-Prüfung, ausgedrückt durch die Ergebnisse der Itemanalyse, ist zurzeit noch nicht konstant. Sie sollte für die entsprechenden Items durch die dargestellten Maßnahmen zur Steigerung der Testgüte für künftige OSCE-Prüfungen verbessert werden, obwohl einige Stationen bereits eine gute Reliabilität zeigen. Die Reliabilität zählt folglich teilweise noch zu den Schwächen dieser Durchführung. Die Objektivität wird durch die standardisierten Checklisten und Rahmbedingungen umgesetzt.

Das Ergebnis dieser Studie zeigt, dass durch eine gute Raterschulung und standardisierte Prüfungsbedingungen eine hohe Interrater-Übereinstimmung erreicht werden kann, sodass die zahnärztlichen/ ärztlichen Rater in formativen Prüfungsformaten durch studentische Rater ersetzt werden können (Chenot und Ehrhardt 2003; Chisnall et al. 2015). Es ergaben sich keine Unterschiede für die Bewertung der verschiedenen Rater. Dies ist im Hinblick auf die Intention einer OSCE-Prüfung wünschenswert, um gleiche Prüfungsbedingungen für die Studierenden zu schaffen.

4.3.4 Weiterführende Fragestellung

Die Untersuchung der Rater-Übereinstimmung zwischen den Fremdurteilen der Studierenden und der Ärzte/ Zahnärzte ist die Basis für einen möglichen Einsatz studentischer Rater in kommenden OSCE-Prüfungen. Die hohen und signifikanten Zusammenhänge zwischen den Bewertungen durch die studentischen und die zahnärztlichen und ärztlichen Rater wurden bereits ausführlich dargestellt. Studentische Rater dürfen allerdings gemäß der Prüfungsrichtlinien der UMG nicht in summativ konstruierten OSCE-Prüfungen zum Einsatz kommen. Zur Durchführung einer formativen OSCE-Prüfung stellt der Einsatz studentischer Rater jedoch eine Möglichkeit dar, ärztliche personelle Ressourcen zu schonen.

Eine formative OSCE-Prüfung dient dazu, den Studierenden Defizite in ihrem aktuellen Leistungsstand aufzuzeigen und ihnen diese unter Rückmeldung in Form eines konstruktiven Feedbacks vor Augen zu führen. Sie ist primär eine Lehr-OSCE und keine Prüfungs-OSCE (Chenot und Ehrhardt 2003; Chisnall et al. 2015). Eine summative OSCE-Prüfung verfolgt hingegen das Ziel, eine Entscheidung über die Eignung der Studierenden zu treffen und über das Vorhandenseins notwendiger klinischer Kompetenzen zum Bestehen eines durch das Curriculum definierten Abschnitts der zahnärztlich-chirurgischen Lehre (Nikendei und Jünger 2006). Vor dem Hintergrund der erhobenen Daten und der daraus ermittelten Ergebnisse sind studentische Rater unter der Voraussetzung, dass sie vorab an einer ausführlichen Schulung teilnehmen, zur Bewertung der Studierenden einsetzbar. Welchen Zweck erfüllt also der Einsatz studentischer Rater in der OSCE und was sind Vor- und Nachteile der Konzepte?

4.3.5 Diskussion der Fragestellung vor dem Hintergrund bisheriger Publikationen

Auf der einen Seite steht die Durchführung einer formativen OSCE mit studentischer Unterstützung. Diese Art der Prüfung erleichtert eines der grundsätzlichen Probleme, die die Durchführung einer OSCE-Prüfung im Vergleich zu schriftlichen Prüfungen mit sich bringt: Die Etablierung einer OSCE-Prüfung ist verglichen mit schriftlichen Prüfungen ein zeitaufwendiges, personalintensives und kostspieliges Vorhaben, das eine gute Organisation und Strukturiertheit erfordert (Chenot et al. 2007; Zaric und Belfield 2015; Simmenroth-Nayda et al. 2014). Die Möglichkeit zur Schonung personeller Ressourcen mithilfe von studentischer Unterstützung ist ein Argument für die Durchführung einer formativen OSCE-Prüfung. Diese Variante gestaltet sich sowohl kostengünstiger, als auch logistisch einfacher zu lösen (Rolita et al. 2009).

Auch das Konzept der formativen OSCE-Prüfung mit sofortigem Feedback in Anschluss an die Prüfung zeigt gute Ergebnisse und wird in der Literatur auch als „OSCE with immediate feedback (OSCE IF)“ bezeichnet (Zaric und Belfield 2015; Rush et al. 2014). Mithilfe dieser individuellen Feedbacks ist es den Studierenden möglich, ihre Stärken und Schwächen aufgezeigt zu bekommen und effektiv an diesen zu arbeiten. Ein kurzes Feedback im Anschluss an die Prüfung „führt nachgewiesener Maßen nachfolgend zu einer Verbesserung der studentischen Fertigkeiten“ (Nikendei und Jünger 2006, S. 3) (Hodder et al. 1989). Es konnte gezeigt werden, dass dies von studentischer Seite aus bereits als hilfreich und motivierend angesehen wird (Stein et al. 2005; Schuebel et al. 2014; Schiekirka und Raupach 2015). Außerdem zeigen Studien, dass sich insbesondere eine zu Beginn des klinischen Studienabschnitts stattfindende OSCE-Prüfung positiv auf die Kompetenz der Studierenden zur späteren Behandlung von Patienten auswirkt (Ratzmann et al. 2012b). Das Prinzip zielt darauf ab, den Studierenden eine Prüfungssituation zu bieten von der sie profitieren ohne ihnen Angst zu bereiten, und die sie anhand des erhaltenen Feedbacks eigenständig reflektieren können. Eine derartige Prüfungssituation soll von studentischer Seite aus als „non-threatening“ (Zaric und Belfield 2015, S. 586), also „harmlos“ wahrgenommen werden. Die Durchführung einer „OSCE IF“, also einer OSCE mit direktem Feedback kann als formative Prüfungsform auch mit einem „Peer-Feedback“ - Feedback durch Kommilitonen - durchgeführt werden (Cushing et al. 2011). Es können theoretisch natürlich sowohl summative als auch formative OSCE-Prüfungen mit einem anschließenden Feedback kombiniert werden. Dennoch lässt es sich nicht leugnen, dass das formative Prüfungskonzept weniger psychischen Druck auf die Prüflinge ausübt, da die Prüfung nicht bestanden werden muss. Es ist möglich, dass sich dieses Bewusstsein positiv auf die Einstellung der Studierenden gegenüber der Prüfung auswirkt, sie ruhiger in die Prüfung hineingehen und sich ihre Aufnahmefähigkeit für konstruktives Feedback erhöht (Allen et al. 1998; Rushton 2005).

Dieser Effekt kann jedoch auch unerwünscht sein und dazu führen, dass die Studierenden die Prüfung nicht ernst genug nehmen und der Lernaufwand, verglichen mit summativen Prüfungen, vernachlässigt wird. Es konnte bereits gezeigt werden, dass eine formative Prüfung weniger Anreize für Studierende gibt intensiv zu lernen. Demnach setzen summative Prüfungsformen den Hauptanreiz für studentisches Lernverhalten und führen nachweislich zu größeren Lernerfolgen als formative Prüfungen (Raupach et al. 2013).

Aus psychologischer Perspektive betrachtet könnte sich die Bewertung von Studierenden durch Studierende problematisch gestalten, da sich die Studierenden, besonders in einer derart kleinen Fakultät wie der Zahnmedizinischen Fakultät der UMG nahezu alle kennen. Es kann also sowohl für die studentischen Rater schwierig werden, die Prüflinge objektiv zu bewerten, als auch für die Prüflinge unangenehm sein, durch bekannte Kommilitonen bewertet zu werden (Lurie et al. 2006b). Beide Varianten ergeben diese Problematik. Insbesondere in diesem Kontext sind die Ratereinflüsse der verschiedenen Ratergruppe zu hinterfragen (Lurie et al. 2006a). Studien zu diesem Thema zeigen, dass die Mehrheit (64%) der Studierenden von einer gleichartigen Bewertung durch studentische sowie durch ärztliche Rater ausgeht. Es gibt aber auch Studierende, wenn auch minderheitlich vertreten, die denken, dass sie durch ihre Kommilitonen besser oder schlechter bewertet werden würden (Chenot et al. 2007). Dieselbe Studie zeigt auch, dass die formative Bewertung durch studentische Rater qualitativ verbessert werden kann, indem sie an der Lehre dieser Aufgaben beteiligt sind: „[…] senior students who are involved in the teaching of the skills can reliably assess their peers at OSCE stations […]“ (Chenot et al. 2007, S. 1036). Die Integration von studentischen Tutoren in das Lehrkonzept, was auch als „peer teaching“ bezeichnet wird, stellt eine Möglichkeit dar, den Lernerfolg nachhaltig zu verbessern (Ogden et al. 2000; Heylings und Stefani 1997; Rudy et al. 2001).

Auch die Durchführung einer summativen OSCE-Prüfung mit ärztlichen Ratern zeigt Nachteile. Das Argument der schwierigen Gestaltung durch mangelnde personelle Ressourcen wurde bereits dargestellt. Zusätzlich stellt sich dann die Frage, wie mit dem Fall einer nicht-bestandenen Prüfung umgegangen wird. Dazu gibt es zwei mögliche Ansätze, die abhängig von der gewählten Prüfungsform der OSCE sind. Einerseits kann die OSCE als kompensatorische Form durchgeführt werden. Das bedeutet, dass die Gesamtnote für die OSCE-Prüfung über alle Stationen gemittelt wird und der Studierende ein schlechtes Ergebnis an einer Station mit einem guten Ergebnis an einer anderen Station ausgleichen kann. Bei nicht-bestandener Gesamtleistung muss folglich aber auch die gesamte OSCE wiederholt werden. Wird hingegen die non-kompensatorische Form gewählt, so muss an jeder einzelnen Station die vorgegebene Bestehensgrenze erreicht werden (Nikendei und Jünger 2006). Die zuletzt beschriebene Prüfungsform gestaltet sich für eine summative OSCE-Prüfung insofern schwieriger, da die Studierenden auch nur die Stationen wiederholen müssen, die nicht bestanden wurden. Dies ist allerdings in einer laufenden OSCE-Prüfung nicht praktikabel (Nikendei und Jünger 2006) und würde einen gesonderten, neuen Prüfungstag für die „Wiederholer“ erfordern,

an denen dann die jeweiligen nicht bestandenen Stationen wiederholt werden müssten. Dies ist mit hohem Aufwand verbunden. Das Format kompensatorisch durchzuführen ist im Hinblick auf die personellen und logistischen Ressourcen der UMG praktikabler. So können die Studierenden mit nicht-bestandener Prüfung die OSCE im Ganzen nachholen. Es ist außerdem möglich eine schlechte Station auszugleichen. Ein weiterer Nachteil der non-kompensatorischen Form ist, dass eine mangelhaft konzipierte Station schlecht differenzierte Ergebnisse zwischen den Leistungen der Studierenden liefert. Außerdem sollte es auch für leistungsstarke Studierende legitim sein, Schwächen im Bereich einer einzelnen Station zu zeigen, ohne dass dies gleich in einer nicht-bestandenen Prüfungsstation resultiert. (Nikendei und Jünger 2006; Chesser et al. 2004; Friedman 2000).

Die Durchführung einer OSCE-Prüfung stellt eine Neuheit in der Zahnklinik der UMG dar und ist den Studierenden noch fremd. Es ist nicht zu ignorieren, dass die Durchführung einer summativen Prüfungsform und das damit in Zusammenhang stehende „Durchfallen“ durch die Prüfung abschreckend und demotivierend auf die Studierenden wirken könnte und ihnen die Möglichkeiten einer OSCE-Prüfung negativ darstellt. Unter anderem aus diesem Grund wurde an der UMG mit einer formativen OSCE-Prüfung begonnen. Nicht-bestandene Prüflinge wurden mündlich individuell nachgeprüft und so die Prüfung langsam in Richtung summativ entwickelt.

Für die Durchführung einer summativen OSCE-Prüfung mit ärztlichen Ratern spricht, dass der Aspekt der Befangenheit aufgrund von persönlichen Differenzen oder Vorlieben zwischen den Studierenden untereinander in dieser Prüfungssituation minimiert ist. Die Bewertung ist neutraler und z. T. wahrscheinlich unbefangener.

Ein entscheidender Vorteil des Einsatzes ärztlicher Rater ist jedoch die Durchführbarkeit einer Vergabe von Feedback im Anschluss an die summative Prüfung. Die Leistungen der Studierenden werden dadurch sofort in einem kurzen persönlichen Gespräch kommuniziert, ohne dabei auf den Lernerfolg einer summativen Prüfungsform, wie es Raupach beschreibt, zu verzichten (Raupach et al. 2013). Ein Feedback in einer summativ durchgeführten Prüfung ist hilfreich, es dreht sich jedoch bei dieser Prüfungsform vorrangig um die Entscheidung über notwendige Kenntnisse und Leistungen zum Bestehen eines definierten Curriculums. Das Feedback erfüllt hier also nicht die klassische Aufgabe, „den Studenten ein konstruktives Feedback über ihren aktuellen Wissens- und Könnensstand geben zu können“ (Nikendei und Jünger 2006), um dann in einer summativen Prüfung erfolgreicher zu bestehen. Das erhaltene Feedback bietet den Studierenden Ansatzpunkte zur Verbesserung der eigenen Leistungen in den praktischen Kursen und kommenden Prüfungen. Ein Nachteil dieser Methode ist, dass ein Studierender sein Feedback an die Kommilitonen der anderen Durchgänge kommunizieren kann, insbesondere wenn er einen Fehler begangen hat, der zum Durchfallen durch die OSCE-Prüfung geführt hat. Aus dieser Hilfe resultieren ungleiche Prüfungsbedingungen (Höfer et al. 2013). Beispielsweise an der Station Lokalanästhesie konnte gezeigt werden, dass sich die palatinale Anästhesie vor Zahnextraktion im Vergleich zum ersten

Durchgang in der Haupt-OSCE deutlich verbessert hat. Es ist allerdings auch bekannt, dass bereits die Durchführung der OSCE alleine, auch ohne Feedback, aufgrund des Test-Effekts zur Verbesserung der studentischen Fähigkeiten und Fertigkeiten führt (Roediger und Karpicke 2006).

4.3.6 Fazit

Die Vor- und Nachteile der beiden Variationen zur Umsetzung der OSCE-Prüfung wurden in diesem Abschnitt gegenübergestellt. Was ist nun also die am meisten geeignete Variante für die Prüfung an der UMG?

Das Hauptproblem bleibt: Um den Ansprüchen an eine hochwertige, summative OSCE-Prüfung mit ärztlichen Ratern gerecht zu werden, sind mindestens zehn, besser zwölf Stationen notwendig (Chenot und Ehrhardt 2003; Nikendei und Jünger 2006), was eine hohe personelle Belastung mit sich bringt. Bis zur Etablierung weiterer sechs Stationen, um diese Anforderungen an eine summative Prüfung zu erfüllen, dürfen zu deren Pilotierung studentische Rater anstelle von ärztlichen Ratern eingesetzt werden. Die Analyse der Interraterreliabilität zeigt, dass durch eine gute Schulung der studentischen Rater eine ebenso hohe Reliabilität erreicht werden kann wie durch ärztliche Rater.

5. Zusammenfassung

Zur Beurteilung zahnärztlich-chirurgischer Fähigkeiten und Fertigkeiten im Rahmen einer objektiven strukturierten klinischen Prüfung (OSCE) können geschulte studentische und ärztliche Rater gleichermaßen eingesetzt werden. Es konnte gezeigt werden, dass mithilfe einer guten Schulung eine hohe Interrater-Übereinstimmung zwischen ärztlichen und studentischen Ratern in der Zahnmedizin erreicht werden kann. Diese ist somit nicht allein von dem Grad der Ausbildung abhängig, sondern kann durch adäquate und ausführliche vorab stattfindende Schulungen und standardisierte Prüfungsbedingungen positiv beeinflusst werden. Es ergeben sich für die studentischen Rater mit niedrigerem Ausbildungsgrad vergleichbare Ergebnisse wie für die ärztlichen Rater mit hohem Ausbildungsgrad und sehr hohe und signifikante Zusammenhänge zwischen beiden Urteilen. Damit bestätigen wir die Ergebnisse einer Studie für medizinische Prüfungen im OSCE-Format (Chenot et al. 2007). Zwar können studentische Rater nicht in summativen Prüfungen eingesetzt werden, zur Pilotierung weiterer OSCE-Stationen jedoch bietet der Einsatz geschulter studentischer Rater eine Möglichkeit, personelle Ressourcen zu schonen. Dies konnte durch diese Studie eindeutig gezeigt werden.

Aus den Ergebnissen der Itemanalyse resultiert die Weiterentwicklung einiger Stationen zur Steigerung der Testgüte, um den künftigen Einsatz der OSCE als summative Prüfung in der ZMK-Klinik der UMG vorzubereiten. Andere Stationen zeigen bereits jetzt eine gute Reliabilität, jedoch werden mindestens sechs weitere Stationen bis zur Etablierung einer hochwertigen, summativen Prüfung zu pilotieren sein. Bis dahin wird die Pilotierung weiterer OSCE-Stationen mit studentischer Unterstützung durchgeführt und eine Weiterentwicklung des Prüfungsformats an der UMG verfolgt. Es ist möglich, dass durch eine Intensivierung des Schulungskonzepts in Zukunft noch höhere Zusammenhänge und Reliabilitäten zwischen den beiden Fremdurteilen erreicht werden können. Zur Schonung personeller Ressourcen wäre es ebenfalls interessant zu ermitteln, inwieweit diese hohen Zusammenhänge auf die Bewertung durch anderes zahnärztliches Personal zu übertragen sind. Ein Projekt zur Bestimmung dieser Zusammenhänge ist zukünftig geplant. Es könnte anschließend möglich sein, zusätzlich zu den studentischen Ratern als Ersatz für die ärztlichen Rater im Rahmen einer formativen OSCE-Prüfung auch geschultes zahnärztliches Personal einzusetzen.

Weitere Projekte zum Thema OSCE und deren Optimierung sind zukünftig geplant. Die OSCE-Prüfung stellt eine valide, objektive und reliable Prüfungsform dar, um klinisch-praktische Fertigkeiten aus dem Arbeitsfeld der zahnärztlichen Chirurgie mit hohem Praxisbezug zu evaluieren.

6. Literaturverzeichnis

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

Bärlocher F: Biostatistik. 2., unveränderte Auflage; Georg Thieme Verlag KG, Stuttgart 2008

Bortz - Lehrbuch der Statistik. Für Sozialwissenschaftler. 2., vollständig neu bearbeitete und erweiterte Auflage; Springer Medizin Verlag, Berlin 1985

Bortz - Statistik für Human- und Sozialwissenschaftler. 6., vollständig überarbeitete und aktualisierte Auflage; Springer Medizin Verlag, Heidelberg 2005

Bortz - Forschungsmethoden und Evaluation. 2., vollständig überarbeitete Auflage; Springer Medizin Verlag, Berlin 1997

Boulet J, De Champlain A, McKinley D (2003): Setting defensible performance standards on OSCEs and standardized patient examinations. Med Teach 25 (3), 245–249

Chenot J-F, Ehrhardt M (2003): Objective structured clinical examination (OSCE) in der medizinischen Ausbildung: Eine Alternative zur Klausur. Z Allg Med 79 (9), 437–442

Chenot J-F, Simmenroth-Nayda A, Koch A, Fischer T, Scherer M, Emmert B (2007): Can student tutors act as examiners in an objective structured clinical examination?. Med Educ 41 (11), 1032–1038

Chesser A, Laing M, Miedzybrodzka Z, Brittenden J, Heys S (2004): Factor analysis can be a useful standard setting tool in a high stakes OSCE assessment. Med Educ 38 (8), 825–831

Chisnall B, Vince T, Hall S, Tribe R (2015): Evaluation of outcomes of a formative objective structured clinical examination for second-year UK medical students. Int J Med Educ 6, 76–83

Cohen J (1992): A power primer. Psychol Bull 112 (1), 155–159

Cushing A, Abbott S, Lothian D, Hall A, Westwood O (2011): Peer feedback as an aid to learning-what do we want? Feedback. When do we want it? Now!. Med Teach 33 (2), 105-112

Davis MH (2003): OSCE: the Dundee experience. Med Teach 25 (3), 255–261

Frank JR (Hrsg.): TheCanMEDS 2005 physician competency framework. Better standards. Better physicians. Better care. The Royal College of Physicians and Surgeons of Canada, Ottawa 2005

Friedman BD (2000): Standard setting in student assessment. An extended summary of AMEE Medical Education Guide No 18. Med Teach 22, 120-130

Friedman M, Mennin SP (1991): Rethinking critical issues in performance assessment. Acad Med 66 (7), 390–395

Giesler M, Forster J, Biller S, Fabry G (2011): Entwicklung eines Fragebogens zur Erfassung von Kompetenzen in der Medizin: Ergebnisse zur Reliabilität und Validität. GMS Z Med Ausbild 28 (2), 1-15

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

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

Hecker KG, Adams CL, Coe JB (2012): Assessment of first-year veterinary students' communication skills using an objective structured clinical examination: the importance of context. J Vet Med Educ 39 (3), 304–310

Heylings DJ, Stefani LA (1997): Peer assessment feedback marking in a large medical anatomy class. Med Educ 31 (4), 281–286

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, Regehr G, Hanson M, McNaughton N (1998): Validation of an objective structured clinical examination in psychiatry. Acad Med 73 (8), 910–912

Höfer SH, Schuebel F, Sader R, Landes C (2013): Development and implementation of an objective structured clinical examination (OSCE) in CMF-surgery for dental students. J Craniomaxillofac Surg 41 (5), 412–416

Kassam A, Cowan M, Donnon T (2016): An objective structured clinical exam to measure intrinsic CanMEDS roles. Med Educ Online 21 (1), 31085

Kiehl C, Simmenroth-Nayda A, Goerlich Y, Entwistle A, Schiekirka S, Ghadimi BM (2014): Standardized and quality-assured video-recorded examination in undergraduate education: informed consent prior to surgery. J Surg Res 191 (1), 64–73

Landes CA, Hoefer S, Schuebel F, Ballon A, Teiler A, Tran A (2014): Long-term prospective teaching effectivity of practical skills training and a first OSCE in cranio maxillofacial surgery for dental students. J Craniomaxillofac Surg 42 (5), 97-104

Lurie SJ, Nofziger AC, Meldrum S, Mooney C, Epstein RM (2006a): Temporal and group-related trends in peer assessment amongst medical students. Med Educ 40 (9), 840–847

Lurie SJ, Nofziger AC, Meldrum S, Mooney C, Epstein RM (2006b): Effects of rater selection on peer assessment among medical students. Med Educ 40 (11), 1088–1097

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

Möltner A, Schellberg D, Jünger J (2006): Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild 23 (3), 53-63

Napankangas R, Karaharju-Suvanto T, Pyorala E, Harila V, Ollila P, Lahdesmaki R, Lahti S (2014): Can the results of the OSCE predict the results of clinical assessment in dental education?. Eur J Dent Educ 20 (1), 2-8

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

Nikendei C, Jünger J (2006): OSCE - praktische Tipps zur Implementierung einer klinisch-praktischen Prüfung. GMS Z Med Ausbild 23 (3), 47-54

Ogden GR, Green M, Ker JS (2000): Student examiners: 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

Pelka M, Farmand M: Dokumentierte Patientenaufklärung, Basisinformation zum Aufklärungsgespräch, Extraktion. pro Compliance (Thieme Compliance GmbH), Erlangen 2012

Rasch B, Friese M, Hofmann WJ, Naumann E: Quantitative Methoden 1. Einführung in die Statistik für Psychologen und Sozialwissenschaftler. Band 1, 4. Aufl.; Springer-Verlag, Berlin 2014(a)

Rasch B, Friese M, Hofmann WJ, Naumann E: Quantitative Methoden 2. Einführung in die Statistik für Psychologen und Sozialwissenschaftler. Band 2, 4. Aufl.; Springer Verlag, Berlin 2014(b)

Ratzmann A, Wiesmann U, Kordaß B (2012a): Integration of an Objective Structured Clinical Examination (OSCE) into the dental preliminary exams. GMS Z Med Ausbild 29 (1), 9-15

Ratzmann A, Wiesmann U, Kordaß B (2012b): Integration einer OSCE in das zahnmedizinische Physikum. GMS-Z Med Ausbild 29 (1), 42–48

Raupach T, Munscher C, Beissbarth T, Burckhardt G, Pukrop T (2011): Towards outcome-based programme evaluation: using student comparative self-assessments to determine teaching effectiveness. Med Teach 33 (8), 446-453

Raupach T, Brown J, Anders S, Hasenfuss G, Harendza S (2013): Summative assessments are more powerful drivers of student learning than resource intensive teaching formats. BMC Med 11, 61-70

Read EK, Bell C, Rhind S, Hecker KG (2015): The use of global rating scales for OSCEs in veterinary medicine. PloS one 10 (3), 1371-1379

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

Reznick RK, Regehr G, Yee G, Rothman A, Blackmore D, Dauphinee D (1998): Process-rating forms versus task-specific checklists in an OSCE for medical licensure. Medical Council of Canada. Acad Med 73 (10), 97-99

Roediger HL, Karpicke JD (2006): The Power of Testing Memory: Basic Research and Implications for Educational Practice. Perspect Psychol Sci 1 (3), 181–210

Rolita L, Ark TK, Moroz A, Lanyi V, Southwell J, Sutin D (2009): Evaluation of an Elderly Faller by Medical Students and Rehabilitation Residents. J Am Geriatr Soc 57 (4), 709-713

Rudy DW, Fejfar MC, Griffith CH, Wilson JF (2001): Self- and peer assessment in a first-year communication and interviewing course. Eval Health Prof 24 (4), 436–445

Rumsey D, Majetschak B, König H (2010): Statistik. Für Dummies. Die Grundlagen der Statistik mit Spaß erlernen und anwenden. 2. Auflage; WILEY-VCH Verlag, Weinheim 2010

Rush S, Ooms A, Marks-Maran D, Firth T (2014): Students' perceptions of practice assessment in the skills laboratory: an evaluation study of OSCAs with immediate feedback. Nurse Educ Pract 14 (6), 627–634

Rushton A (2005): Formative assessment: a key to deep learning?. Med Teach 27 (6), 509–513

Sachs L, Hedderich J: Angewandte Statistik. Methodensammlung mit R. 13. Auflage; Springer-Verlag, Heidelberg 2009

Schelten A: Grundlagen der Testbeurteilung und Testerstellung. Teststatistik und Testtheorie für Pädagogen und Ausbilder in der Praxis. Quelle & Meyer, Heidelberg 1980

Schiekirka S, Raupach T (2015): A systematic review of factors influencing student ratings in undergraduate medical education course evaluations. BMC Med Educ 15, 30-38

Schiekirka S, Reinhardt D, Heim S, Fabry G, Pukrop T, Anders S, Raupach T (2012): Student perceptions of evaluation in undergraduate medical education: A qualitative study from one medical school. BMC Med Educ 12, 45-52

Schiekirka S, Reinhardt D, Beissbarth T, Anders S, Pukrop T, Raupach T (2013): Estimating learning outcomes from pre- and posttest student self-assessments: a longitudinal study. Acad Med 88 (3), 369–375

Schiekirka S, Anders S, Raupach T (2014): Assessment of two different types of bias affecting the results of outcome-based evaluation in undergraduate medical education. BMC Med Educ 14, 149-157

Schlosser O: Einfühung in die sozialwissenschaftliche Zusammenhangsanalyse. Rowohlt Taschenbuch Verlag, Reinbek bei Hamburg 1976

Schnell R, Hill PB, Esser E: Methoden der empirischen Sozialforschung. 5. Aufl.; R. Oldenbourg Verlag, München 1995

Schuebel F, Höfer SH, Rüsseler M, Walcher F, Sader R, Landes C (2014): Introduction of craniomaxillofacial surgery as a component of medical student training in general surgery. J Oral Maxillofac Surg 72 (11), 2318

Simmenroth-Nayda A, Görlich Y, Wagner M, Müther M, Lohse C, Utte L (2014): Studentische Lehre in der Augenheilkunde. Sind standardisierte praktische Prüfungen sinnvoll? Ophthalmologe 111 (3), 235–240

Stein MR, Parish SJ, Arnsten JH (2005): The OSCE as a formative evaluation tool for substance abuse teaching. Med Educ 39 (5), 529–530

Tudiver F, Rose D, Banks B, Pfortmiller D (2009): Reliability and validity testing of an evidence-based medicine OSCE station. Fam Med 41 (2), 89–91

Weiß C: Basiswissen Medizinische Statistik. 6. Aufl.; Springer-Verlag, Berlin 2013

Wilkinson TJ, Frampton CM, Thompson-Fawcett M, Egan T (2003): Objectivity in objective structured clinical examinations: checklists are no substitute for examiner commitment. Acad Med 78 (2), 219–223

Wirtz M, Caspar F: Beurteilerübereinstimmung und Beurteilerreliabilität. Hogrefe-Verlag, Göttingen 2002

Zaric S, Belfield L A (2015): Objective Structured Clinical Examination (OSCE) with Immediate Feedback in Early (Preclinical) Stages of the Dental Curriculum. Creative Education 6, 585-593

7. Anhang

7.1 Bögen

7.1.1 Einverständniserklärung der Studierenden für Videoaufzeichnung

UNIVERSITÄTSMEDIZIN GÖTTINGEN : UMG

Universitätsmedizin Göttingen, 37099 Göttingen
Klinik für Mund-, Kiefer- und Gesichtschirurgie, Robert-Koch-Str. 40, D-

Zentrum Zahn-, Mund- und Kieferheilkunde
Klinik für Mund-, Kiefer- und Gesichtschirurgie

Ärztliche Leitung
Prof. Dr. Dr. Henning Schliephake
Sekretariat
Susanne Ashrafi
37099 Göttingen
Robert-Koch-Straße 40, 37075 Göttingen
0551/39-8306/-8343 /-5305
Fax 0551/39-12653
se.ki@med.uni-goettingen.de

19/11/2017

Dokumentation und Einverständniserklärung über Videoaufzeichnung

Das Filmmaterial geht in das Eigentum der UMG über und darf die UMG nicht verlassen. Die Verwendung, Archivierung und Aufbewahrung des Filmmaterials auf einem geeigneten Speichermedium in der Klinik für Mund-, Kiefer- und Gesichtschirurgie erfolgt ausschließlich zum Zwecke der Prüfungsauswertung, die Dateien werden entsprechend der Aufbewahrungsfristen für Prüfungen abschließend gelöscht. Ansprechpartnerin für Rückfragen und Einsicht in die Auswertung ist Frau PD. Dr. med. dent Sennhenn-Kirchner Tel.: 0551-39 14022.

Dateiname des Videotestats = Vorname_Nachname:

…….

……..…

ggf. Laufzeiten des auszuwertenden Videoabschnittes:

von

min/sec ……..……..…..bis ….....…….……

…

Student:

Ich erkläre mich einverstanden

……..……..……

Name in leserlicher Druckschrift

Auf dem Video zu sehen ist / sind:

……..……

Datum, Unterschrift ○ Student(in)

7.1.2 Vorlage aus der Humanmedizin der UMG

Universitätsmedizin Göttingen, 37099 Göttingen
Allgemein-, Viszeral- und Kinderchirurgie, Lehrkoordination,

Zentrum 8 Chirurgie
Klinik für Allgemein-, Viszeral- und Kinderchirurgie
Direktor: Univ.-Prof. Dr. med. Michael Ghadimi

37099 Göttingen **Briefpost**
Robert-Koch-Straße 40, 37075 Göttingen **Adresse**
0551 39-8977**Telefon**
0551 39-6109**Fax**

Dokumentation und Einverständniserklärung für Videoaufzeichnung

Das Filmmaterial geht in das Eigentum der UMG über und darf die UMG nicht verlassen. Die Verwendung, Archivierung und Aufbewahrung des Filmmaterials auf einem geeigneten Speichermedium in der Klinik für Allgemein-, Viszeral- und Kinderchirurgie erfolgt ausschließlich zum Zwecke der Prüfungsauswertung, die Dateien werden entsprechend der Aufbewahrungsfristen für Prüfungen abschließend gelöscht. Ansprechpartnerin für Rückfragen und Einsicht in die Auswertung ist Frau Prof. Dr. med. König Tel.: 0551-39 8977.

Dateiname des Videotestats = Gruppennummer und beide Nachnamen:

…….
……..…

ggf. Laufzeiten des auszuwertenden Videoabschnittes:

von
min/sec ……..……..…..bis ….....…….……
…

Student(in) 1:

Ich erkläre mich einverstanden

……..……..……

Name in leserlicher Druckschrift

Auf dem Video zu sehen ist / sind:

……..……

Datum, Unterschrift

○ Student(in) 1

○ Student(in) 2

Student(in) 2:

Ich erkläre mich einverstanden

……..……..……

Name in leserlicher Druckschrift

……..……

Datum, Unterschrift

7.1.3 Abschlussevaluation der Studierenden

Allgemeine Daten						
Matrikelnummer						
Selbsteinschätzung						
Ich habe die OSCE Prüfung	Bestanden	Nicht bestanden	Weiß nicht			
In der Station I (Einwilligung Zahnextraktion) habe ich folgende Note erreicht	1	2	3	4	5	
In der Station II (postoperative Aufklärung) habe ich folgende Note erreicht	1	2	3	4	5	
In der Station III (Lokalanästhesie) habe ich folgende Note erreicht	1	2	3	4	5	
In der Station IV (Zahnextraktion) habe ich folgende Note erreicht	1	2	3	4	5	
In der Station V (Naht) habe ich folgende Note erreicht	1	2	3	4	5	
In der Station VI (BLS) habe ich folgende Note erreicht	1	2	3	4	5	
Prüfungsformat						
OSCE Prüfungen sind meines Erachtens angemessen für das chirurgische Fachgebiet	Trifft genau zu				Trifft gar nicht zu	Enthaltung
Zusätzlich multiple choice Fragestationen würde ich sehr sinnvoll finden						
Kommunikationsstationen mit Schauspielpatienten sind sehr sinnvoll						
So eine OSCE Prüfung ist der reine Stress!						

Eine praktische Prüfung wie diese sind deutlich sinnvoller als MC Prüfungen						
Ich wünsche mir für das Zahnmedizinstudium strukturierte Lehre bezüglich der Patientenkommunikation						
Anmerkungen: Ich mache folgende Vorschläge zur Weiterentwicklung/Verbesserung der Prüfungsform OSCE						

7.1.4 Beurteilung der Tablet-Software/ tOSCE-Software

	Trifft voll zu	Trifft weitgehend zu	Trifft überwiegend zu	Trifft selten zu	Trifft sehr selten zu	Trifft gar nicht zu
Funktionalität der Software						
Die Applikation bietet mir alle Funktionalitäten um die Studierenden effizient zu bewerten						
Die Anwendung bietet mir Erleichterungen bei sich wiederholenden Arbeitsschritten						
Alle Funktionen in der Applikation lassen sich einfach bedienen						
Software-Ergonomie						
Die Führung durch die Applikation ist für mich klar und einfach						
Alle Symbole und Bezeichnungen sind für mich verständlich und sachgerecht gewählt						
Die gewählten Farben und die verwendete Darstellung ist für mich gut erkennbar und unterscheidbar						
Fehlerrobustheit						
Meine Eingaben werden vom System rechtzeitig auf						

Plausibilität geprüft, Fehler werden mir angezeigt						
Die Fehlermeldungen sind für mich verständlich						
Die Anwendung gibt mir ausreichend Hinweise, wie ich Fehler korrigieren kann						
Eingabefehler lassen sich für mich leicht korrigieren						
Gesamtbeurteilung (Platz für Lob und Kritik)						

7.2 Deskriptive Statistik der Pilot-OSCE

7.2.1 Häufigkeitsverteilungen der Checklisten-Mittelwerte

In den Tabellen 17-21 sind die jeweiligen von den Studierenden erreichten Mittelwerte (errechnet aus allen Items der Station) mit zugehöriger Häufigkeit (Anzahl aus der Gesamtheit N aller Studierenden) und die dazugehörige Prozentzahl bei N = 9 Studierender dargestellt.

Tabelle 17: Häufigkeitsverteilung der Station „Aufklärung und Einwilligung Zahnextraktion“

M	Häufigkeit	Prozent
0,29	1	11,1
0,41	1	11,1
0,42	1	11,1
0,63	1	11,1
0,71	1	11,1
0,79	1	11,1
0,83	1	11,1
0,88	1	11,1
0,96	1	11,1
Gesamt	9	100

Tabelle 18: Häufigkeitsverteilung der Station „Lokalanästhesie“

M	Häufigkeit	Prozent
0,46	1	11,1
0,54	3	33,3
0,62	2	22,2
0,77	2	22,2
0,92	1	11,1
Gesamt	9	100

Tabelle 19: Häufigkeitsverteilung der Station „Zahnextraktion“

M	Häufigkeit	Prozent
0,5	1	11,1
0,68	1	11,1
0,75	1	11,1
0,86	4	44,4
0,93	2	22,2
Gesamt	9	100

Tabelle 20: Häufigkeitsverteilung der Station „Naht“

M	Häufigkeit	Prozent
0,42	1	11,1
0,69	1	11,1
0,73	1	11,1
0,77	1	11,1
0,81	1	11,1
0,85	1	11,1
0,88	3	33,3
Gesamt	9	100

Tabelle 21: Häufigkeitsverteilung der Station „Aufklärung post OP“

M	Häufigkeit	Prozent
0,31	1	11,1
0,35	1	11,1
0,42	1	11,1
0,46	1	11,1
0,5	1	11,1
0,58	1	11,1
0,65	1	11,1
0,77	1	11,1
0,81	1	11,1
Gesamt	9	100

7.2.2 Noten nach Umcodierung mit Häufigkeitsverteilung

Tabelle 22: Häufigkeitsverteilung der Noten

In den Zeilen sind hier die erreichbaren Noten von 1-5 dargestellt und in den Spalten die fünf Stationen der OSCE. Daraus ergibt sich die Anzahl an Studierenden von der Gesamtheit N = 9, welche die jeweilige Noten erreichten.

	Aufklärung, Einwilligung Zahnextraktion	LA	ZE	Naht	Aufklärung post OP
0 (Note 5)	0	0	2	0	0
0,25 (Note 4)	1	3	1	1	1
0,5 (Note 3)	2	4	3	2	6
0,75 (Note 2)	3	2	2	6	1
1 (Note 1)	3	0	1	0	1

7.3 Hauptstudie

In den folgenden Tabellen 23-28 sind die statistischen Werte für die jeweiligen Items der Stationen aufgeführt. Die Items sind in codierter Form dargestellt. Der Mittelwert M (Schwierigkeit), die zugehörige Standardabweichung und die Trennschärfe ergeben sich aus der Gesamtheit der teilnehmenden Studierenden (N=36) und der Alpha-Koeffizient bezieht sich auf das Gesamtergebnis für die Station, wenn das jeweilige Item eliminiert werden würde. Die Werte sind linksseitig für die zahnärztlichen Rater und rechtsseitig für die studentischen Rater dargestellt.

7.3.1 Itemanalyse

Tabelle 23: Itemanalyse Station „Aufklärung und Einwilligung Zahnextraktion“

	Zahnarzt				Studentischer Rater			
Item	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item entfällt	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item entfällt
Item 1	0,931	0,212	0,347	0,712	0,903	0,288	0,372	0,728
Item 2	0,861	0,257	0,134	0,729	0,917	0,224	0,343	0,733
Item 3	0,750	0,349	0,459	0,694	0,597	0,312	0,446	0,720
Item 4	0,750	0,305	0,462	0,696	0,903	0,262	0,495	0,718
Item 5	0,778	0,303	0,348	0,709	0,722	0,303	0,337	0,732
Item 6	0,653	0,355	0,434	0,697	0,694	0,275	0,399	0,726
Item 7	0,486	0,422	0,394	0,702	0,736	0,253	0,065	0,756
Item 8	0,639	0,425	0,203	0,731	0,750	0,254	0,347	0,732
Item 9	0,917	0,280	0,192	0,725	0,389	0,494	0,352	0,737
Item 10	0,500	0,507	0,491	0,686	0,611	0,494	0,512	0,708
Item 11	0,278	0,454	0,376	0,705	0,222	0,422	0,460	0,716
Item 12	0,389	0,449	0,484	0,687	0,472	0,430	0,481	0,713
Mittelwert	0,661	0,360	0,360		0,660	0,334	0,384	
Cronbachs Alpha	0,725				0,744			
Cronbachs Alpha, wenn Gesamteindruck ein Item der Skala ist	0,778				0,782			

Korrelation Checklistenmittelwert und Gesamteindruck: Zahnarzt: 0,901 (p = 0,000, N = 36); Studentischer Rater 0,861 (p = 0,000, N = 36)

Tabelle 24: Itemanalyse Station „postoperative Aufklärung“

	Zahnarzt				Studentischer Rater			
Item	M (Schwie-rigkeit)	SD	Trenn-schär-fe	Alpha, wenn Item entfällt	M (Schwie-rigkeit)	SD	Trenn-schärfe	Alpha, wenn Item ent-fällt
Item 1	0,583	0,423	0,181	0,829	0,861	0,283	0,061	0,647
Item 2	0,806	0,322	0,675	0,778	0,875	0,250	0,395	0,593
Item 3	0,736	0,327	0,458	0,797	0,750	0,305	0,314	0,604
Item 4	0,792	0,277	0,472	0,797	0,889	0,211	0,396	0,598
Item 5	0,917	0,224	0,357	0,806	0,778	0,326	0,425	0,580
Item 6	0,764	0,327	0,694	0,776	0,514	0,304	0,273	0,611
Item 7	0,778	0,367	0,615	0,782	0,569	0,271	0,263	0,613
Item 8	0,736	0,304	0,539	0,791	0,639	0,307	0,479	0,571
Item 9	0,444	0,374	0,483	0,795	0,500	0,169	0,384	0,605
Item 10	0,861	0,227	0,325	0,808	0,583	0,327	0,172	0,632
Item 11	0,806	0,300	0,249	0,814	0,694	0,322	0,136	0,638
Item 12	0,583	0,471	0,582	0,786	0,597	0,460	0,247	0,629
Mittelwert	0,734	0,329	0,469		0,687	0,295	0,295	
Cronbachs Alpha	0,811				0,631			
Cronbachs Alpha, wenn Gesamtein-druck ein Item der Skala ist	0,837				0,694			

Korrelation Checklistenmittelwert und Gesamteindruck: Zahnarzt: 0,874 (p = 0,000, N = 36); Studentischer Rater 0,831 (p = 0,000, N = 36)

Tabelle 25: Itemanalyse Station „Lokalanästhesie“

	Zahnarzt				Studentischer Rater			
Item	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item entfällt	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item entfällt
Item 1	0,944	0,232	0,240	0,458	0,940	0,232	0,158	0,305
Item 2	1,000	0,000	0,000	0,484	1,000	0,000	0,000	0,330
Item 3	0,500	0,507	0,403	0,389	0,470	0,506	0,052	0,334
Item 4	0,444	0,504	0,015	0,508	0,280	0,454	0,131	0,300
Item 5	0,667	0,478	0,296	0,426	0,530	0,506	0,249	0,246
Item 6	0,417	0,500	0,299	0,423	0,390	0,494	0,164	0,286
Item 7	0,833	0,378	0,066	0,485	0,890	0,319	-0,193	0,387
Item 8	0,750	0,439	0,279	0,433	0,860	0,351	0,125	0,305
Item 9	0,889	0,319	-0,151	0,521	0,860	0,351	-0,011	0,345
Item 10	0,694	0,467	0,016	0,504	0,720	0,454	0,177	0,282
Item 11	0,278	0,454	-0,058	0,521	0,500	0,507	-0,234	0,446
Item 12	0,750	0,439	0,350	0,414	0,780	0,422	0,485	0,162
Item13	0,750	0,439	0,211	0,452	0,75	0,439	0,257	0,252
Item 14	0,556	0,504	0,345	0,409	0,640	0,487	0,122	0,304
Mittelwert	0,677	0,404	0,165		0,686	0,394	0,106	
Cronbachs Alpha	0,481				0,328			
Cronbachs Alpha, wenn Gesamteindruck ein Item der Skala ist	0,560				0,431			

Korrelation Checklistenmittelwert und Gesamteindruck:

Zahnarzt: 0,827 (p = 0,000, N = 36)

Studentischer Rater 0,802 (p = 0,000, N = 36)

Tabelle 26: Itemanalyse Station „Zahnextraktion“

	Zahnarzt				Studentischer Rater			
Item	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item entfällt	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item entfällt
Item 1	1,000	0,000	0,000	0,533	1,000	0,000	0,000	0,637
Item 2	1,000	0,000	0,000	0,533	1,000	0,000	0,000	0,637
Item 3	0,556	0,504	0,357	0,466	0,556	0,504	0,332	0,605
Item 4	0,583	0,387	0,091	0,534	0,653	0,375	0,106	0,641
Item 5	0,444	0,392	0,451	0,452	0,569	0,417	0,594	0,556
Item 6	0,389	0,494	0,295	0,485	0,417	0,500	0,091	0,654
Item 7	0,528	0,506	-0,116	0,599	0,417	0,500	0,247	0,623
Item 8	0,806	0,401	0,169	0,518	0,833	0,378	0,377	0,599
Item 9	0,944	0,232	0,181	0,517	0,944	0,232	0,210	0,625
Item 10	0,708	0,385	0,333	0,481	0,819	0,297	0,382	0,603
Item 11	0,653	0,393	-0,008	0,556	0,514	0,470	0,293	0,613
Item 12	0,750	0,305	0,152	0,520	0,708	0,302	0,148	0,632
	0,847	0,312	0,125	0,525	0,903	0,288	0,321	0,611
Item 13	0,389	0,242	0,117	0,525	0,681	0,271	0,209	0,625
Item 14	0,806	0,401	0,502	0,437	0,806	0,401	0,343	0,604
Item 15	0,972	0,167	0,391	0,500	0,972	0,167	0,221	0,626
Mittelwert	0,711	0,320	0,190		0,737	0,319	0,242	
Cronbachs Alpha	0,530				0,635			
Cronbachs Alpha, wenn Gesamteindruck ein Item der Skala ist	0,594				0,695			

Korrelation Checklistenmittelwert und Gesamteindruck: Zahnarzt: 0,741 (p = 0,000, N = 36); Studentischer Rater 0,831 (p = 0,000, N = 36)

Tabelle 27: Itemanalyse Station „Naht“

	Zahnarzt				Studentischer Rater			
Item	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item entfällt	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item entfällt
Item 1	0,917	0,189	0,398	0,591	0,542	0,325	0,266	0,612
Item 2	0,236	0,348	0,452	0,562	0,458	0,403	0,292	0,606
Item 3	0,806	0,275	0,145	0,621	0,500	0,431	0,278	0,610
Item 4	0,681	0,320	0,171	0,619	0,764	0,327	0,165	0,627
Item 5	0,861	0,283	0,221	0,609	0,833	0,338	0,157	0,629
Item 6	0,861	0,283	0,221	0,609	0,903	0,201	0,026	0,639
Item 7	0,778	0,422	0,496	0,544	0,778	0,422	0,200	0,625
Item 8	0,389	0,361	0,340	0,586	0,444	0,444	0,530	0,552
Item 9	0,542	0,385	0,203	0,617	0,583	0,455	0,399	0,583
Item 10	0,778	0,422	0,293	0,598	0,639	0,487	0,294	0,608
Item 11	0,944	0,232	0,145	0,619	0,944	0,232	0,097	0,633
Item 12	1,000	0,000	0,000	0,626	1,000	0,000	0,000	0,635
Item 13	0,875	0,220	0,224	0,609	0,833	0,316	0,578	0,562
Mittelwert	0,744	0,288	0,255		0,709	0,337	0,252	
Cronbachs Alpha	0,621				0,630			
Cronbachs Alpha, wenn Gesamteindruck ein Item der Skala ist	0,675				0,691			

Korrelation Checklistenmittelwert und Gesamteindruck:

Zahnarzt: 0,768 (p = 0,000, N = 36)

Studentischer Rater 0,836 (p = 0,000, N = 36)

Tabelle 28: Itemanalyse Station „BLS“

	Zahnarzt				Studentischer Rater			
Item	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item entfällt	M (Schwierigkeit)	SD	Trennschärfe	Alpha, wenn Item entfällt
Item 1	0,000	0,000	0,000	0,282	0,000	0,000	0,000	0,310
Item 2	0,940	0,232	0,340	0,204	0,970	0,167	0,196	0,278
Item 3	0,640	0,487	0,177	0,217	0,690	0,467	0,153	0,263
Item 4	0,670	0,478	0,106	0,255	0,640	0,487	0,000	0,347
Item 5	0,610	0,494	0,046	0,287	0,890	0,319	0,062	0,303
Item 6	0,780	0,422	0,145	0,238	0,890	0,319	0,000	0,325
Item 7	0,860	0,351	-0,001	0,296	0,890	0,319	0,000	0,325
Item 8	0,530	0,506	0,153	0,229	0,440	0,504	0,078	0,306
Item 9	0,780	0,422	0,099	0,259	0,780	0,422	0,304	0,187
Item 10	0,830	0,378	-0,256	0,391	0,830	0,378	0,098	0,290
Item 11	0,470	0,506	0,199	0,202	0,610	0,494	0,289	0,179
Item 12	0,940	0,232	0,175	0,246	1,000	0,000	0,000	0,310
Mittelwert	0,671	0,376	0,099		0,719	0,323	0,098	
Cronbachs Alpha	0,280				0,307			
Cronbachs Alpha, wenn Gesamteindruck ein Item der Skala ist	0,428				0,431			

Korrelation Checklistenmittelwert und Gesamteindruck:

Zahnarzt: 0,827 (p = 0,000, N = 36)

Studentischer Rater 0,829 (p = 0,000, N = 36)

7.3.2 Deskriptive Statistik

Die Abkürzungen in den Spalten sind wie folgt zu verstehen: Chl = Checklisten-Mittelwerte; Note = Globalnote; A = zahnärztliche/ärztliche Rater; S = studentische Rater. In den Zeilen sind hier die Stichprobengröße N, der Checklisten-Mittelwert mit zugehöriger Standardabweichung SD sowie die Parameter der Verteilungseigenschaften Schiefe und Kurtosis und das Minimum (Min) und Maximum (Max) dargestellt. Rot markiert sind die Werte für Schiefe und Kurtosis, die sich außerhalb einer Normalverteilung befinden. Annähernd normalverteilt sind Verteilungen, wenn Schiefe und Kurtosis nahe Null liegen (akzeptabler Bereich: Schiefe zwischen -0,5 und +0,5; Kurtosis zwischen -1 und 1).

Tabelle 29: Deskriptive Statistik der Station POA

	POA_chl_A	POA_Note_A	POA_A_gesamt	POA_chl_S	POA_Note_S	POA_S_gesamt	POA_Note_Selbst
N	36	36	36	36	36	36	34
Mittelwert	0,73	2,33	0,71	0,69	2,61	0,66	2,24
SD	0,19	0,83	0,19	0,13	0,77	0,15	0,55
Schiefe	-0,81	0,25	-0,54	0,02	-0,80	0,47	2,36
Kurtosis	0,06	-0,29	-0,36	-0,96	0,27	-0,64	4,77
Minimum	0,25	1	0,33	0,46	1	0,43	2
Maximum	1	4	1	1	4	1	4

Tabelle 30: Deskriptive Statistik der Station LAE

	LAE_chl_A	LAE_Note_A	LAE_A_gesamt	LAE_chl_S	LAE_Note_S	LAE_S_gesamt	LAE_Note_Selbst
N	36	36	36	36	36	36	34
Mittelwert	0,68	2,89	0,63	0,69	2,86	0,64	2,59
SD	0,15	0,89	0,17	0,13	0,80	0,15	0,66
Schiefe	0,08	-0,03	0,15	-0,09	-0,45	0,06	0,68
Kurtosis	-0,93	0,10	-0,66	-0,62	0,08	-0,35	-0,49
Minimum	0,430	1	0,300	0,430	1	0,380	2
Maximum	1	5	1	1	4	1	4

Tabelle 31: Deskriptive Statistik der Station ZE

	ZE_chl_A	ZE_Note_A	ZE_A_gesamt	ZE_chl_S	ZE_Note_S	ZE_S_gesamt	ZE_Note_Selbst
N	36	36	36	36	36	36	31
Mittelwert	0,71	3,08	0,64	0,74	2,72	0,69	2,26
SD	0,13	0,84	0,14	0,14	1,06	0,17	0,82
Schiefe	-0,54	0,14	-0,19	-0,30	0,29	-0,23	0,27
Kurtosis	0,67	0,73	0,88	-0,18	-0,35	-0,49	-0,21
Minimum	0,380	1	0,260	0,410	1	0,350	1
Maximum	1	5	1	1	5	1	4

Tabelle 32: Deskriptive Statistik der Station Naht

	Naht_chl_A	Naht_Note_A	Naht_A_gesamt	Naht_chl_S	Naht_Note_S	Naht_S_gesamt	Naht_Note_Selbst
N	36	36	36	36	36	36	33
Mittelwert	0,74	2,89	0,68	0,71	2,50	0,68	2,91
SD	0,13	0,71	0,14	0,15	0,91	0,17	0,95
Schiefe	-0,12	-0,35	0,09	-0,30	-0,24	-0,12	0,43
Kurtosis	-0,56	0,35	-0,39	-0,46	-0,69	-0,61	-0,14
Minimum	0,460	1	0,400	0,350	1	0,320	1
Maximum	1	4	1	1	4	1	5

Tabelle 33: Deskriptive Statistik der Station BLS

	BLS_chl_A	BLS_Note_A	BLS_A_gesamt	BLS_chl_S	BLS_Note_S	BLS_S_gesamt	BLS_Note_Selbst
N	36	36	36	36	36	36	33
Mittelwert	0,67	2,81	0,63	0,72	2,25	0,71	2,70
SD	0,14	0,95	0,16	0,12	0,73	0,14	0,73
Schiefe	-0,05	-0,22	0,14	-0,53	0,03	-0,26	0,55
Kurtosis	0,43	0,08	0,68	-0,13	-0,27	-0,45	-0,88
Minimum	0,330	1	0,230	0,420	1	0,430	2
Maximum	1	5	1	1	4	1	4

7.4 Intra-Klassen-Korrelation

Tabelle 34: Die Intra-Klassen-Korrelation der ärztlichen und studentischen Rater

Dargestellt sind hier die einzelnen Stationen der OSCE (Zeilen) und die Intra-Klassen-Korrelation zwischen den ärztlichen und studentischen Ratern, die Mittelwerte/durchschnittlichen Schwierigkeiten, zugehörigen Trennschärfen und Cronbachs Alpha für die ärztlichen Rater bzw. die studentischen Rater (Spalten).

OSCE mit allen Stationen							
Station	ICC ärztlicher Rater – studentischer Rater	durchschnittliche Schwierigkeit (ärztlicher Rater)	durchschnittliche Trennschärfe (ärztlicher Rater)	Cronbachs Alpha (ärztlicher Rater)	Durchschnittliche Schwierigkeit (studentischer Rater)	Durchschnittliche Trennschärfe (studentischer Rater)	Cronbachs Alpha (studentischer Rater)
AEZ	0,682	0,661	0,360	0,725	0,660	0,384	0,744
POA	0,495	0,734	0,469	0,811	0,687	0,295	0,631
LA	0,794	0,677	0,165	0,481	0,686	0,106	0,328
ZE	0,737	0,711	0,190	0,530	0,737	0,242	0,635
Naht	0,705	0,744	0,255	0,621	0,709	0,252	0,630
BLS	0,653	0,671	0,099	0,280	0,719	0,098	0,307
Mittelwert	0,678	0,700	0,256	0,575	0,700	0,230	0,546
Gesamt-OSCE	0,888						

Danksagung

Ich möchte mich auf diesem Wege herzlich bei Frau Prof. Dr. König und Frau Dr. Simmenroth-Nayda aus der Medizinischen Fakultät der Universitätsmedizin Göttingen bedanken, die mir durch ihre Erfahrungen mit der Durchführung der OSCE-Prüfungen stets unterstützend zur Seite standen, mir eine Hospitation ermöglichten und seither ihr OSCE-Equipment für unsere zahnmedizinischen Prüfungen zur Verfügung stellen.

Außerdem bedanke ich mich bei Marcel Notbohm und seinen Kollegen aus dem Institut für Medizinische Informatik der UMG, die uns nicht nur die technischen Voraussetzungen zur Durchführung dieser OSCE schufen, sondern auch die elektronische Übertragung der Daten managten. Besonderer Dank gilt in dieser Hinsicht auch Yvonne Görlich aus der Abteilung für Statistik, die selbst in das größte Zahlenkonstrukt Ordnung brachte.

Ich bedanke mich außerdem bei allen zahnärztlichen und ärztlichen Ratern, die sich für diesen Tag aus ihrem regulären Dienst zurückzogen und uns bei der Bewertung der Studierenden halfen. Ohne diese Hilfe wäre es unmöglich gewesen, die OSCE durchzuführen.

Der größte Dank gilt allerdings Frau Dr. Sennhenn-Kirchner, die nicht nur für mich jederzeit als Doktormutter verfügbar und offen für Fragen, Anmerkungen und Hilfestellungen war, sondern auch generell allen Studierenden während der OSCE-Prüfung mit Rat und Tat zur Seite stand. Die erfolgreiche Durchführung dieser OSCE ist Resultat ihres unermüdlichen Engagements für die Studierenden der Zahnmedizin in Göttingen und die Weiterentwicklung der Lehre, und es war mir eine Freude, Teil dieses Projekts werden zu dürfen. Für ihre Hilfe und Unterstützungen auf dem Weg zur Vollendung dieser Arbeit gilt ihr von Herzen mein ewiger Dank!