Aus dem Universitätsklinikum Münster
Poliklinik für Zahnersatzkunde des
Zentrums für Zahn-, Mund- und Kieferheilkunde
-Direktor: Univ.-Prof. Dr. Dr. F. Bollmann-

Zur inter- und intraindividuellen Reliabilität der Beurteilung vorklinischer Zahnersatzarbeiten mittels Checkliste

INAUGURAL-DISSERTATION
zur
Erlangung des doctor medicinae dentium
der Medizinischen Fakultät der
Westfälischen Wilhelms-Universität Münster

vorgelegt von
Schiffler, Christian Philipp
aus Duisburg

2007

Gedruckt mit Genehmigung der
Medizinischen Fakultät der
Westfälischen Wilhelms-Universität Münster

Dekan:	Univ.-Prof. Dr. V. Arolt
1. Berichterstatter:	Univ.-Prof. Dr. P. Scheutzel
2. Berichterstatter:	Prof. Dr. E. Schäfer
Tag der mündlichen Prüfung:	27.08.2007

Aus dem Universitätsklinikum Münster
Poliklinik für Zahnersatzkunde des
Zentrums für Zahn-, Mund- und Kieferheilkunde
-Direktor: Univ.-Prof. Dr. Dr. F. Bollmann-
Referentin: Univ.-Prof. Dr. P. Scheutzel
Koreferent: Prof. Dr. E. Schäfer

Zusammenfassung

Zur inter– und intraindividuellen Reliabilität der Beurteilung vorklinischer Zahnersatzarbeiten mittels Checkliste

Schiffler, Christian Philipp

In der Medizin ergibt sich grundsätzlich die Notwendigkeit der Leistungsbewertung bzw. der Beurteilung klinischer Situationen in Hinsicht auf die Ergebnissqualität. Klassische Gütekriterien für die Reliabilität einer Bewertung, auch bei studentischen Arbeiten in der Zahnmedizin, sind die Objektivität (interindividuelle Reliabilität) und die Reproduzierbarkeit (intraindividuelle Reliabilität).

Ziel der vorliegenden Untersuchung ist es, das Ausmaß der in der Person des Bewertenden begründeten inter- und intrapersonellen Variabilität (Grad der Objektivität und Reproduzierbarkeit) bei der Benotung praktischer vorklinischer Studentenarbeiten am Phantomkopf (hier: Kunststoffverblendbrücke 24 - 26) festzustellen. In diesem Zusammenhang soll untersucht werden, ob ein statistisch signifikanter Unterschied zwischen verschiedenen Bewertergruppen mit unterschiedlicher Berufserfahrung (vorklinische Studenten, klinische Studenten, vorklinische Zahnärzte, klinische Zahnärzte) besteht und inwieweit die Selbst- bzw. Fremdeinschätzung von Studierenden des vorklinischen Phantomkurses verlässlich ist. Hierzu wurden 30 anonymisierte Kunststoffverblendbrücken des Phantomkurses II durch vier Bewertergruppen in zwei zeitlich getrennten Durchgängen bewertet. Als Grundlage dieser Bewertung dienten, wie in der zahnmedizinischen Ausbildung allgemein üblich, Checklisten. Hinter den zu beurteilenden Teilaspekten wurde die nach der „glance and grade“ Methode (subjektive Einstufung nach Inaugenscheinnahme) ermittelte Note eingetragen. Für die Benotung waren auf der Checkliste keine Vorgaben in Form von Bewertungskriterien gemacht.

Die Ergebnisse zeigen deutliche Unterschiede in der Notenvergabe. Die Eigenbewertung durch die vorklin. Stud. selber und die Bewertung durch die klin. ZÄ ergaben deutlich bessere Noten als die Bewertung durch die klin. Stud. und die vorklin. ZÄ. Diese Notendifferenzen waren bei der Bewertung von Arbeiten des oberen Notendrittels am stärksten, wogegen Arbeiten des unteren Leistungsdrittels durch die vor - klin. und klin. ZÄ sowie die klin. Stud. weitgehend einheitlich bewertet wurden. Die geringste interindividuelle Urteilsübereinstimmung bestand innerhalb der Gruppe der vorklinischen Zahnärzte wo je nach beurteiltem Teilaspekt der Brücke nur Korrelationen von $r_p = 0{,}11$ bis $r_p = 0{,}56$ erreicht wurden und somit nur eine ungenügende Objektivität gegeben war. Was die Beurteilung von Teilaspekten betrifft, so ergab sich für die Bereiche „Approximalkontakt“ und „Okklusion“ bei den Bewertern nur eine geringe interindividuelle Urteilsübereinstimmung, wogegen in Bezug auf die Bewertung der Passgenauigkeit (Schaukeln, ja/nein) in allen Gruppen die höchste Übereinstimmung erreicht wurde. Bezüglich der intrapersonellen Urteilskonkordanz, d.h. der Reproduzierbarkeit bei wiederholter Bewertung, erreichten alle Bewerter eine gute Übereinstimmung ($r_p = 0{,}72 – 0{,}83$).

Aus den vorliegenden Ergebnissen wird geschlossen, dass auch unter Benutzung einer Checkliste, die zwar die zu bewertenden Teilaspekte einer Arbeit auflistet, jedoch keine Vorgaben zu deren Bewertungskriterien macht, durchaus eine reliable Bewertung vorklinischer Brückenarbeiten möglich ist, wie die Ergebnisse der klin. ZÄ und klin. Stud. zeigen, andererseits aber ebenso mit einer starken Urteilsvariabilität, d. h. nur geringer Objektivität gerechnet werden muss, wie es bei den vorkl. ZÄ und den vorkl. Stud. der Fall war. Insofern sollte die bestehensrelevante Benotung vorklinischer Phantom-Arbeiten möglichst durch mehrere Bewerter unterschiedlicher Berufserfahrung erfolgen (z. B. Kursleiter und Kursassistenten). Desweiteren sollte für die Zukunft durch die Entwicklung und den Einsatz von detaillierten Bewertungsbögen versucht werden, die Objektivität der Benotung weiter zu steigern und insbesondere auch eine verlässlichere Basis für die Eigenbewertung der Studierenden zu schaffen.

Tag der mündlichen Prüfung: 27.08.2007

Diese Dissertation ist meinen Eltern gewidmet

Inhaltsverzeichnis

1 Einleitung

1.1 Einführung

In der Medizin ergibt sich grundsätzlich die Notwendigkeit der Leistungsbewertung bzw. der Beurteilung klinischer Situationen in Praxis und Ausbildung. Speziell restaurative Fächer der Zahnmedizin verlangen täglich Entscheidungen, ob die Ergebnisqualität das Einsetzen einer Restauration gestattet oder nicht.
Klassische Gütekriterien für die Reliabilität einer Bewertung sind die Objektivität (interindividuelle Reliabilität) und die Reproduzierbarkeit (intraindividuelle Reliabilität). Das Kriterium Objektivität (interindividuelle Reliabilität) gibt an, wieweit das Ergebnis eines Tests abhängig davon ist, wer die Untersuchung durchgeführt hat, wobei zwischen der Durchführungsobjektivität, Auswertungsobjektivität und Interpretationsobjektivität unterschieden wird. Eine Methode zur Überprüfung der Objektivität lautet: Ein Test ist objektiv, wenn mehrere Untersucher mit dem gleichen Test bei der gleichen Population die gleiche Häufigkeitsverteilung der Testwerte registrieren. Bei der Reproduzierbarkeit (intraindividuelle Reliabilität) geht es um die Wiederholbarkeit und damit Zuverlässigkeit der Datenerhebung in Anlehnung an das Messkonzept in der Physik, wobei man davon ausgeht, dass die zu messenden Merkmale über die Zeit stabil sind. Eine möglichst hohe Reliabilität ist die Grundlage für eine valide Bewertung. Die Validität (Gültigkeit) als drittes Qualitätsmerkmal jeder Beurteilung gibt an, inwieweit ein Test/Messinstrument tatsächlich das misst, was zu messen ist. Hierbei wird zwischen Inhalts-, Konstrukt- und Kriterienbezogener Validierung unterschieden.
Was die Beurteilung vorklinischer studentischer Arbeiten in der Zahnmedizin betrifft, ist in erster Linie die Reliabilität (Zuverlässigkeit) der Bewertung/Benotung von Bedeutung. Als Grundlage für die Bewertung vorklinischer Zahnersatzkundearbeiten am Phantompatienten oder im Labor dient in der Regel eine Checkliste (Aufzählung von Merkmalen, die einen Gegenstand umfassend beschreiben), die alle zu untersuchenden Punkte enthält und dazu dienen soll, keinen relevanten Teilaspekt zu übersehen oder zu vergessen. Diese Checklisten enthalten allerdings keine Bewertungskriterien, sondern le-

diglich eine Aufzählung der zu bewertenden Teilaspekte. Die Bewertung an sich erfolgt in der Regel durch „glance and grade“ (Inaugenscheinnahme der vorgegebenen Merkmale, allein anhand von aus Sicht des Bewertenden allgemeingültigen Qualitätskriterien).

Die wenigen bisher erfolgten Untersuchungen zur Reliabilität einer solchen Benotung geben allerdings kaum Hinweise darauf, wie hoch die Objektivität (interindividuelle Reliabilität) und die Reproduzierbarkeit (intraindividuelle Reliabilität) bei der Beurteilung der Ergebnisqualität vorklinischer Zahnersatzarbeiten ist. Hier existiert bisher lediglich eine Pilotstudie von *Türp et al* [30], die darauf hindeutet, dass die Variabilität der Benotung vorklinischer Zahnersatzarbeiten sehr hoch ist. Vor diesem Hintergrund ergibt sich die dringende Notwendigkeit weiterführender Untersuchungen zur Klärung des Ausmaßes der inter- und intraindividuellen Reliabilität bei der Beurteilung vorklinischer Zahnersatzarbeiten mittels Checkliste.

1.2 Forschungsfrage und Ziel der Untersuchung

Ziel der vorliegenden Untersuchung ist es, das Ausmaß der in der Person des Bewertenden begründeten inter- und intrapersonellen Variabilität bei der Benotung praktischer vorklinischer Studentenarbeiten am Phantomkopf (hier: Kunststoffverblendbrücke 24 - 26) festzustellen. Hierbei ergeben sich die folgenden Forschungsfragen:

1. Welchen Einfluss hat die Erfahrung des Beurteilenden auf die Objektivität und die Reproduzierbarkeit der Beurteilung vorklinischer studentischer Arbeiten (Kunststoffverblendbrücken) des Phantomkurses?

In diesem Zusammenhang soll untersucht werden, ob ein statistisch signifikanter Unterschied zwischen verschiedenen Bewertergruppen mit unterschiedlicher Berufserfahrung (vorklinische Studenten, klinische Studenten, vorklinische Zahnärzte, klinische Zahnärzte) besteht.

2. Wie verlässlich ist die Selbst- bzw. Fremdeinschätzung von Studierenden des vorklinischen Phantomkurses bei der Beurteilung des von ihnen selbst bzw. einem Kommilitonen gefertigten festsitzenden Zahnersatzes (Kunststoffverblendbrücke)?

In diesem Zusammenhang soll untersucht werden, wie sich die Selbsteinschätzung der Studierenden von der Bewertung anderer unterscheidet und ob Unterschiede auch in Grenzbereichen vorliegen.

2 Literaturübersicht

2.1 Überblick über bisherige Studien zur Reliabilität der Bewertung vorklinischer studentischer Phantomarbeiten

Bei einer systematischen Literaturrecherche zum Thema der Bewertungsreliabilität in den Datenbanken *Medline*, *Cochrane Library* und *Current Content Medicine*, ergänzt durch eine Handsuche nach dem Schneeballprinzip, konnten von 1950 bis 2006 nur 31 Studien identifiziert werden. Hierbei stammt lediglich eine Veröffentlichung (*Türp et al*. [30]) aus dem deutschsprachigen Raum, alle übrigen sind im englischsprachigen Raum erschienen.

Gegenstand der Untersuchungen waren prothetische und konservierende Arbeiten. Im prothetischen Bereich beschäftigten sich *Türp et al.* [30] mit der Bewertung von Verblendkronen und partiellen Prothesen, *Feil et al.* [6] ebenfalls mit Verblendkronen, *Gaines et al.* [8] mit Wachsmodellationen und *Bedi et al.* [2] mit Kronenpräparationen. Alle übrigen Studien untersuchten die Bewertung konservierender Leistungen, wie zum Beispiel Klasse II-Präparationen, Amalgamfüllungen, Inlays und Wurzelkanalbehandlungen. Alle Studien fanden nur eine geringe Reliabilität bei der Bewertung der studentischen Arbeiten mit Reliabilitätskoeffizienten von $r_p = 0{,}11$ bis $r_p = 0{,}86$, bzw. ICC = 0,23 bis ICC = 0,68.

Als allgemeine Schlussfolgerungen zur Erklärung der unterschiedlichen Reliabilität wurden mehrere Ansätze formuliert. So scheint zum Beispiel die Art der Beurteilung Einfluss zu haben, d. h., ob - wie bisher allgemein üblich - nur anhand einer Checkliste (Aufzählung von Merkmalen, die einen Gegenstand umfassend beschreiben) oder mittels genau definierter Bewertungskriterien bewertet wird. Außerdem ist das Maß der Variabilität der Bewertung abhängig von der Art der Definition benutzter Kriterien. Es soll mit genau definierten Kriterien bzw. beispielhaft erklärten Kriterien verringert werden [4, 6, 8, 32]. In diesem Zusammenhang wurde auch der Einsatz von Bewertungsbögen und der Einfluss der Anzahl der Bewertungsunterpunkte diskutiert [6, 13, 14]. Zusätzlich scheint auch die Detaillierung vorhandener Bewertungskriterien eine Rolle zu

spielen [6, 8, 13, 14, 23]. Neben diesen, die Art der Bewertung betreffenden Kriterien, scheint allerdings auch die Person des Bewerters an sich (z. B. Berufserfahrung, Kalibrierung) Einfluss auf die Bewertung zu haben [1, 2, 4, 7, 9, 13, 14, 22, 23, 30]

Autor	**bewertete Arbeit**	**interind. Reliabilität (Objektivität)**	**intraind. Reliabilität (Reproduzierbarkeit)**
Bedi et al. (1987)	Klasse I-, II-, III-Kronen-Präparationen	52,4-84,8%; k 0,29-0,78	
Feil et al. (1982)	Klasse II-AgAm, VMK-Kronen	ICC= 0,68-0,8	ICC= 0,53-0,68
Fuller et al. (1972)	Klasse II- Präparationen	r_s= 0,2-0,56	r_s= 0,47-0,83
Gaines et al. (1974)	Wachsmodellationen	ICC=0,26-0,56	
Goepferd et al. (1980)	Klasse II-Präparationen	ICC= 0,3-0,47	r_p= 0,62-0,68
Hinkelmann et al. (1973)	Kavitätenpräparation, Füllungen	46,7%-84%	
Houpt et al. (1973)	Klasse II-Präparationen	r_{Fin}= 0,61-0,75	r_p= 0,36-0,63
Lilley et al., (1968)	Amalgamfüllungen	r_p= 0,11-0,72	r_p= 0,51-0,63
Meetz et al. (1988)	technical skills	r= 0,62-0,83	
Robertello et al. (1997)	Amalgamfüllungen	61-70%	83-92%
Salvendy et al. (1976)	Amalgamfüllungen	κ_t= 0,36-0,64	0,48-0,73
Türp et al., (2002)	Verblendbrücken	ICC= 0,61	
Vann et al. (1983)	Klasse-II-Kavitäten	ICC= 0,32-0,49	r_p= 0,46-0,86

Tab. 1 *Zusammenstellung bisheriger Untersuchungs-Ergebnisse zur Reliabilität der Bewertung vorklinischer Studentenarbeiten. (Hier wurden nur Studien berücksichtigt, in denen die Reliabilität mittels Reliabilitätskoeffizient metrisch erfasst wurde: r_s = Rangkorrelationskoeffizient nach Spearman, r_p = Produkt-Moment-Korrelationskoeffizient nach Pearson, ICC = Intraclass Correlation Coefficient, κ_t = Kendalls Tau).*

2.2 Einflussfaktoren auf die Reliabilität der Bewertung vorklinischer Studentenarbeiten

Trotz der Versuche, Bewertungsfehler zu minimieren, ergaben die verschiedenen Studien eine inkonstante geschätzte Reliabilität. Die Werte hierfür lagen in Bereichen von r_s = 0,2 bis 0,83 bzw. ICC = 0,23 bis 0,68 [5, 6, 7, 8, 12, 16, 24, 26]. Tab. 1 zeigt die von den verschiedenen Autoren ermittelten Werte. Diese starken Schwankungen in den Reliabilitätskoeffizienten erlauben kaum einen Rückschluss auf Verbesserungsmöglichkeiten [8]. Somit ist das Maß des Einflusses verschiedener Faktoren schwierig zu bestimmen und vorhandene bzw. ermittelte Werte sind in gewisser Weise als fragwürdig zu betrachten.

2.2.1 Beurteilungskriterien: Reliabilitätssteigerung durch exakte Definition?

Das Ausmaß des Bewertungsirrtums hängt in erster Linie von der Subjektivität des Bewerters ab. Daher wird eine Verringerung dieser Subjektivität angestrebt.
Es wurden zahlreiche Manipulationen an Art und Kriterien der Bewertung vorgenommen. Diese veränderten Bewertungskriterien sollten dem Bewerter die Möglichkeit geben, problemlos eine objektive Meinung mit ihnen auszudrücken. Einen Versuch, dies umzusetzen, stellt die genaue Definition der Kriterien dar [2, 4, 6, 7, 8, 14, 26, 30]. Die Notenwahl soll durch Kriterien vereinfacht werden, welche die einzelnen Qualitätsniveaus genau beschreiben. Der Bewerter soll durch die Inspektion einer Arbeit und den Vergleich des Gesehenen mit den ihm zur Verfügung stehenden beschriebenen Arten der Ausführung zur objektiven Bewertung geleitet werden. Daraus resultiert die direkte Assoziation einer bestimmten Note mit erlaubten bzw. begangenen Fehlern. Das Notenniveau wird dadurch anschaulich [2, 6, 7, 14, 26, 33].
Ebenfalls wurde versucht, über eine Aufschlüsselung der Notenniveaus in verschiedene Unterpunkte, den Weg zur Note auszuweisen. Bei dieser Methode werden Kriterien für jeden Unterpunkt bewertet, und je nach Art der Erfüllung wird die Bewertung des jeweiligen Punktes in die eine oder andere Richtung gelenkt.

Verschiedene Autoren haben unterschiedliche Erfolge mit Manipulationen an den Bewertungskriterien erzielt. Eine Übersicht gibt Tab. 2. *Fuller et al.* [7] erreichten weder durch Bewertungsbögen mit vorgegebenen aufgeschlüsselten Kriterien noch durch Training eine gewünschte Reliabilitätssteigerung bei der Bewertung von Klasse-II-Präparationen. Für die interindividuelle Urteilskonkordanz lagen die ICCs bei 0,4 für die Bewertung ohne und bei 0,38 für die Bewertung mit Bogen. Ähnliches beschrieben *Vann et al*. [33]. Sie untersuchten ebenfalls die Bewertung von Klasse-II-Kavitäten bei Einsatz verschiedener Bewertungsmethoden. Weder die globale noch die analytische oder die Checklist-Only-Methode erbrachte eine Steigerung der intraindividuellen oder interindividuellen Reliabilität. Es wurden allerdings auch positive Ergebnisse beschrieben. *Bedi et al*. [2] berichteten über deutlich höhere Kappa-Koeffizienten bei der Nutzung definierter Kriterien im Vergleich zur klassischen globalen Methode. Vor allem unerfahrene Bewerter konnten von diesem System profitieren und so ihre interindividuelle Urteilskonkordanz bei der Bewertung von Klasse–I-, Klasse–II-, Klasse-III– und Kronen-Präparationen steigern. Auch *Gaines et al*. [8] erzielten positive Ergebnisse durch den Einsatz eines strukturierten Bewertungsbogens. Die intraindividuelle Reliabilität bei der Bewertung von Wachsmodellationen konnte von einem ICC = 0,26 auf 0,56 angehoben werden. Zusätzlich lag eine deutliche Steigerung der interindividuellen Urteilskonkordanz vor. Die Untersuchung von *Robertello et al*. [26] zeigte ebenfalls, dass ein Bewertungsbogen mit definierten Kriterien die Reliabilität der Bewertung, in diesem Fall von Amalgamfüllungen, steigern kann

Autor	**bewertete Arbeit**	**ohne Bogen intra-/interindiv. Urteilskonkordanz**	**mit Bogen intra-/interindiv. Urteilskonkordanz**
Bedi et al. (1987)	Klasse-I-, -II-, -III-, Kronen-Präparationen	- / k 0,32	- / k=0,72
Dhuru et al. (1978)	Klasse-II-Präparationen	- / ICC= 0,52	- / ICC= 0,65
Feil et al. (1982)	Klasse-II-AgAm, VMK-Kronen	- / -	ICC=0,6 / 0,74
Fuller et al. (1972)	Klasse-II-Präparationen	r_s= 0,72 / ICC=0,4	- / ICC=0,38
Gaines et al. (1974)	Wachsmodellationen	ICC= 0,26 / -	ICC= 0,56 / -
Goepferd et al. (1980)	Klasse-II-Präparationen	r_p= 0,62 / ICC= 0,3	r_p= 0,68 / ICC=0,47
Houpt et al. (1973)	Klasse-II-Präparationen	r_p= 0,63 / r_{Fin}= 0,61	r_p=0,36 / r_{Fin}=0,75
Natkin et al. (1967)	Wurzelkanalbehandlungen	- / Notenabw. 4,16	- / Notenabw.3,34
Robertello et al. (1997)	Amalgamfüllungen	83 / 61%	92 / 70%
Türp et al. (2002)	Verblendbrücken	ICC= 0,61	
Vann et al. (1983)	Klasse-II-Kavitäten	r_p= 0,75 / ICC= 0,34	r_p= 0,73 / ICC= 0,33#

Tab. 2 *Reliabilitätssteigerung durch Einsatz eines definierten Bewertungssystems.*

2.2.2 Bewertungssystem: Verbesserung durch ein einfaches System mit weniger Notenstufen, genauen Definitionen und ausführlicher Beschreibung?

Ein anderer in verschiedenen Untersuchungen aufgegriffener Aspekt ist die Anzahl von Bewertungspunkten. Es wurde vermutet, dass eine höhere Anzahl möglicher Notenpunkte die Übereinstimmung zwischen den Bewertern verringert. Neben *Hinkelmann und Long* [13] beschrieben *Houpt und Kress* [14] eine größere Übereinstimmung bei Nutzung eines Zwei-Punkte-Systems im Vergleich zu einem System mit fünf möglichen Punkten. Letzteres versprach hingegen einen besseren Lehreffekt für die Ausführung der bewerteten Klasse-II-Päparationen. Es muss bei dieser Argumentation bzw. bei Manipulationen am Punktsystem in dieser Richtung berücksichtigt werden, dass allein die Struktur (mehr Punkte = mehr Möglichkeiten) höhere Abweichungen auftreten lässt und somit die Änderung nicht an der Subjektivität des Bewerters ansetzt.
Gaines et al. [8], *Hinkelmann und Long* [13] und *Ryge et al*. [27] verbesserten in ihren Studien die Spezifität der Punkte durch eine klare Definition von Kriterien. Sie berichteten von großen Übereinstimmungen der Bewerter. Zu ähnlichen Ergebnissen kamen auch andere Autoren, wie Tab. 3 zeigt.

Autor	**bewertete Arbeit**	**unspez.. Notenkriteien intra/interindiv. Urteilskonkordanz**	**def. Notenkriterien intra/interindiv. Urteilskonkordanz**
Goepferd et al. (1980)	Klasse II-Präparationen	r_p= 0,62 / ICC= 0,3	r_p= 0,68 / ICC=0,47
Houpt et al. (1973)	Klasse II-Präparationen	r_p= 0,71 / r_{Fin}= 0,54	r_p=0,56 / r_{Fin}=0,83
Natkin et al. (1967)	Wurzelkanalbehandlungen	- / Notenabw. 4,16	- / Notenabw.3,34
Robertello et al. (1997)	Amalgamfüllungen	83 / 61%	92 / 70%
Vann et al. (1983)	Klasse-II-Kavitäten	r_p= 0,75 / ICC= 0,34	r_p= 0,73 / ICC= 0,33

Tab. 3 *Abhängigkeit der Reliabilität vom Vorhandensein oder Fehlen definierter Kriterien.*

2.2.3 Bewerter: Kalibrierung und Berufserfahrung

Entscheidender Einflussfaktor für die Reliabilität einer Bewertung ist der Bewerter selbst. Viele Autoren versuchten deshalb, Einfluss auf diesen Unsicherheitsfaktor zu nehmen. So lag das Bestreben mehrerer Versuche darin, eine Kalibrierung verschiedener Bewerter zu erzielen. *Houpt und Kress* [14] verglichen die vergebenen Noten anschließend mit einer Expertenmeinung als Goldstandard dieser Benotung, so dass die Bewerter ihre eigenen Noten überdenken konnten. In der nächsten Bewertungsrunde sollte dann die Expertenmeinung bei der Bewertung ähnlicher Situationen erinnerlich sein und die Bewertung beeinflussen können. *Fullers* Kalibrierungsversuch bestand aus einem zweistündigem Training mit anschließender Diskussionsrunde [7]. Auf diese Weise sollte den Teilnehmern die richtige Art der Notenvergabe für das einzelne Leistungsniveau nahe gebracht werden. Weder *Houpt und Kress* [14] noch *Fuller et al*. [7] erzielten den gewünschten Erfolg. Hingegen stellten *Natkin und Guild* [23] eine Steigerung der Reliabilität durch Diskussionstreffen fest. Auch *Abou-Rass* [1] konnte durch Trainingssitzungen die interindividuelle Reliabilität steigern. *Dhuru et al*. [4] griffen das Thema des Einflusses der Bewertererfahrung auf. In ihrer Studie bewerteten sowohl erfahrene als auch unerfahrene Bewerter eine Reihe von Klasse-II-Präparationen in verschiedenen Durchgängen mit und ohne Checkliste. Insgesamt war die geringste Reliabilität bei den unerfahrenen Bewertern ohne Checkliste, die größte Reliabilität bei den erfahrenen Bewertern mit Checkliste festzustellen. Auffällig erschien, dass die Reliabilität der unerfahrenen Bewerter mit Checkliste nur geringfügig höher war als die Reliabilität der erfahrenen Bewerter ohne Checkliste. Diese Ergebnisse zeigten einen positiven Einfluss der größeren Erfahrung auf das Maß an Reliabilität.

Die unten abgebildete Tab. 4 gibt eine Übersicht der Veröffentlichungen, welche sich besonders mit dem Faktor Berufserfahrung befassen. Kamen zusätzliche Hilfsmittel, wie z. B. Trainingssitzungen oder Checklisten, zum Einsatz, so ist dies vermerkt.

Autor		**Bewertererfahrung**		
		gering	**mittel**	**viel**
	Hilfsmittel	**nein/ja**	**nein/ja**	**nein/ja**
Bedi et al. (1987)		k= 0,33 / -		k= 0,39 / -
Dhuru et al. (1978)	Checkliste	ICC 0,47 / 0,54		ICC0,52 / 0,65
Goepferd et al. (1980)		ICC= 0,3-0,47 / -		r_p= 0,62-0,68 / -
Hinkelman et al. (1973)	Training	63,4% / 49,2%		56% / 68%
Meetz et al., (1988)		r= 0,62-0,83 / -		
Natkin et al. (1967)		durchschnittliche Notenabweichung		
		4,16 / -	3,69 / -	3,34 / -
Türp et al. (2002)		ICC= 0,61		

Tab. 4 *Einfluss der Bewertererfahrung auf die Reliabilität.*

3 Material und Methode

3.1 Allgemeiner Untersuchungsablauf

Die vorliegende Studie wurde im Sommersemester 2003 nach Ende des Phantomkurses der Zahnersatzkunde II in der Poliklinik für zahnärztliche Prothetik des Universitätsklinikums Münster durchgeführt. Zu diesem Zeitpunkt standen die Kursnoten bereits fest und konnten als Referenz (Goldstandard) dienen. Die Bewerter kannten bei der Abgabe ihrer Wertungen diese Noten nicht. Dies sollte eine Verfälschung der eigenen Benotung verhindern.

Die vorklinischen Studenten (Gruppe I) bewerteten jeweils ihre eigene Arbeit und die zufällig ausgewählte Arbeit eines Kommilitonen. So wurde jede Arbeit doppelt eingeschätzt. Für die Untersuchung durch die anderen Bewerter-Gruppen (II-IV) wurden die 30 Modelle mit den bewerteten Arbeiten in zufälliger Reihenfolge angeordnet und von den Teilnehmern bewertet. Die Bewertung fand unter konstanten Lichtverhältnissen statt. Als Untersuchungshilfe stand den Teilnehmern eine zahnärztliche Sonde zur Verfügung. In einem zeitlichen Abstand von mehr als 24 Stunden nach dem ersten Durchgang wurden die Arbeiten erneut bewertet. Hierzu erfolgte die Modellanordnung in veränderter Reihenfolge, um die Beeinflussung des Urteils durch Erinnerungen an den vorherigen Durchgang zu minimieren. Der Vergleich der Ergebnisse des ersten mit denen des zweiten Durchgangs liefert das Maß an intraindividueller Reliabilität der Bewerter. Vergleicht man die Urteile der einzelnen Gruppen untereinander, so erhält man die Werte für die interindividuelle Reliabilität.

Die Benotungen wurden auf einer Checkliste vermerkt (siehe 3.4). Anschließend erfolgte die Bewertung der studentischen Arbeiten mit Hilfe eines strukturierten Bewertungsbogens und ebenfalls nachfolgender Wiederholung. Die dabei ermittelten Ergebnisse sind Bestandteil der Studie von *Kellersmann* [17]. Abb.1 gibt den zeitlichen Ablauf der Untersuchung schematisch wieder.

Kunststoffverblendbrücke 24 - 26

n = 30

1. Durchgang

Bewertung durch

- vorklinische Studenten selbst
- zufällig ausgewählten Kommilitonen

Bewertungsmethode

- Checkliste (vorliegende Untersuchung)
- strukturierter Bewertungsbogen (*Kellersmann* [17])

2. Durchgang

Bewertung durch

- klinische Studenten (n = 2, 8. Semester)
- vorklinische Zahnärzte (n = 2, 1,5 bzw. 2 Jahre Berufserfahrung)
- klinische Zahnärzte (n = 2, 4 bzw. 6 Jahre Berufserfahrung)

Bewertungsmethode

- Checkliste, mit Wiederholung (vorliegende Untersuchung)
- Strukturierter Bewertungsbogen, mit Wiederholung (*Kellersmann* [17])

Abb. 1 *Ablauf der Untersuchung.*

3.2 Beurteilte Arbeiten

Für die Studie wurden 30 repräsentative Arbeiten des Phantom-Kurses II der Zahnersatzkunde des Sommersemesters 2003 ausgewählt, wobei die Auswahl so erfolgte, dass das gesamte Notenspektrum gleichmäßig vertreten war (bezogen auf die Kursnote). Die ausgewählten Arbeiten wurden durch eine Nummerierung von 1 bis 30 anonymisiert.
Bei den Arbeiten handelte es sich pro Student um eine Kunststoffverblendbrücke 24 – 26 (Abb. 2). Die Präparationen wurden an Kunststoffzähnen der Firma KaVo, welche in KaVo-Modellen fixiert waren, durchgeführt. Die Modelle waren vollbezahnt, mit Ausnahme des Zahnes 25. Die entsprechende Alveole war mit Silikon geschlossen. Um den Zahnersatz anzufertigen, wurde eine Korrekturabformung mit Silikon genommen und ein Sägemodell aus Superhartgips hergestellt. Auf diesem wurden die Gerüste aufgewachst und anschließend im Gießverfahren in Metall überführt. Einzig bei der Kunststoffverblendung erhielten die Studenten die professionelle Hilfe eines Zahntechnikers. Wie unter Patientenbedingungen erfolgte die letzte Anpassung der „laborfertigen“ Arbeiten am Phantommodell.

Abb. 2 *Phantommodell mit Kunststoffverblendbrücke 24 – 26.*

3.3 Bewerter

	Bewerter	**Tätigkeit**	**Berufserfahrung, Zeitpunkt der Untersuchung**
Gruppe I	Stud.	Studierende des vorklinischen Studienabschnittes; Kursteilnehmer des Phantom-Kurs II der Zahnersatzkunde	klinisch keine, zwei technische Kurse absolviert
Gruppe II	Klin. ZA 1 Klin. ZA 2	Betreuung der klinischen Behandlungskurse in der zahnärztl. Prothetik	Examen vor 3 bzw. 7 Jahren; Studentenbetreuung über 7 bzw. 14 Semester (vorklinische und klinische Kurse)
Gruppe III	Vorkl. ZA 1 Vorkl. ZA 2	Betreuung der vorklin. Studentenkurse	Examen vor 8 bzw. 10 Monaten, Studentenbetreuung über 1,5 bzw. 2 Semester (vorklinische Kurse)
Gruppe IV	Kl. Stud. 1 Kl. Stud. 2	Studierende des klinischen Studienabschnittes: Kursteilnehmer des Behandlungskurs II der Zahnersatzkunde	vorklinische Ausbildung abgeschlossen; drei klinische Behandlungskurse (prothetische und konservierende Zahnheilkunde) absolviert

Tab. 5 *Charakterisierung der in der Studie eingesetzten Bewerter.*

3.4 Bewertungsbögen

Die Bewerter erhielten Checklisten, auf denen acht typische zu beurteilende Merkmale für festsitzenden Zahnersatz tabellarisch aufgelistet waren.

Diese waren:

- Randschluss
- Passgenauigkeit (Schaukeln, Drehen, Friktion)
- Approximalkontakt
- Brückenglied
- Okklusion/Höhe
- Okklusalflächengestaltung/Zahnform
- technische Verarbeitung
- Ästhetik (Form, Farbe).

Abb. 3a bis e veranschaulichen diese Punkte. Abschließend wurden eine Gesamtbeurteilung unter klinischen Gesichtspunkten und eine Note festgesetzt. Für die Beurteilung waren auf der Checkliste keine Vorgaben durch Bewertungskriterien gemacht. Vielmehr sollte jeder Bewerter nach der „glance and grade“ Methode entsprechend seinem subjektiven Qualitätsempfinden, welches sich an dem im Unterricht dargestellten allgemeinen Qualitätsstandards orientieren sollte, entscheiden. Für jedes Qualitätsmerkmal wurde ein Kreuz entweder bei „optimal“, „gut“, „noch akzeptabel“ oder bei „nicht akzeptabel“ vermerkt. Die Gesamtnote der Brücke (1 bis 6) wurde auf Basis der Bewertung der oben genannten Unterpunkte ermittelt, wobei keine Vorgaben zur Gewichtung gemacht wurden.

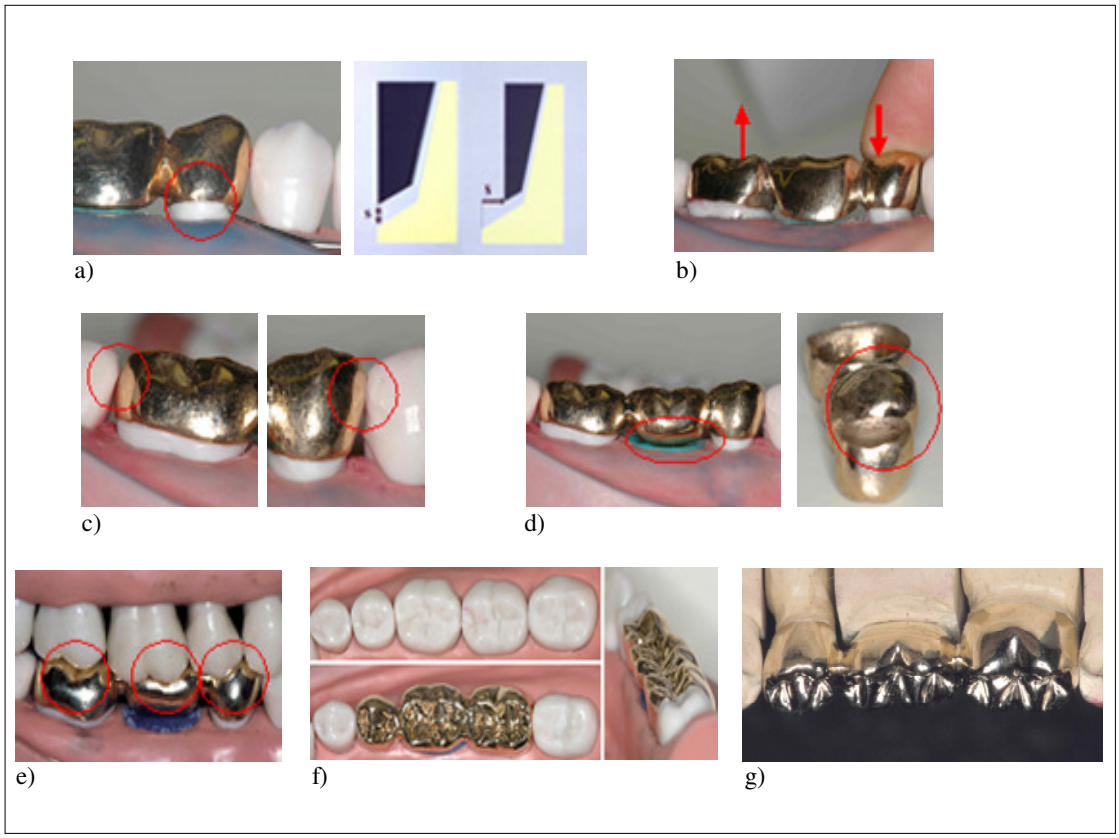

Abb. 3 *Darstellung der Kriterien: a) Randschluss, b) Passgenauigkeit, c) Approximalkontakt, d) Brückenglied, e) Okklusion/Höhe, f) Okklusalflächengestaltung/Zahnform, g) Technische Verarbeitung/Ästhetik.*

3.5 Statistische Auswertung

Die Auswertung der Daten erfolgte als explorative Datenanalyse unter dem Betriebssystem Microsoft Windows XP®. Alle erhobenen Daten wurden zunächst mittels Excel® für Windows XP in Tabellenform erfasst. Die statistische Auswertung geschah anschließend mit dem Programm-Paket SPSS® 13.0 für Windows XP (SPSS Inc. Chicago USA). Die graphische Darstellung der Ergebnisse erfolgte mit Hilfe des Programms Microsoft Office Power Point 2003®.

Als statistische Maßzahlen zur Beschreibung des Datenmaterials wurden der arithmetische Mittelwert und die jeweilige Standardabweichung in den Stichproben bestimmt. Die Normalverteilung der Mittelwerte wurde mit Hilfe des Kolmogorov-Sirnov-Anpassungs-Tests [28], die Varianzhomogenität mit dem Levene-Test [28] geprüft. Zur Beurteilung der Objektivität (interindividuelle Reliabilität) wurde die durchschnittliche Notendifferenz zwischen den Bewertern derselben Bewertungsgruppen ermittelt. Inwieweit die zwischen den Notendifferenzen bestehenden Unterschiede statistisch signifikant sind, wurde mittels t-Test für verbundene Stichproben überprüft [28]. Von einem statistisch signifikanten Ereignis wird gesprochen, wenn die Irrtumswahrscheinlichkeit, der Fehler 1. Art (α) unter 5 % liegt ($p < 0,05$).

Als Maß für die interindividuelle Übereinstimmung der Bewertungen der jeweiligen Bewerter derselben Gruppe wurde der Produkt-Moment-Korrelationskoeffizient nach Pearson (r_p) bestimmt [28]. Der Korrelationskoeffizient r_p kann Werte von -1 bis 1 annehmen, wobei für den praktischen Nutzen Werte von 0 bis 1 interessant sind. Ein Wert von 0,00 gibt an, dass keine Übereinstimmung vorliegt, während der Wert 1,00 eine perfekte Übereinstimmung darstellt. Bei Werten von $r_p \leq 0,4$ wird von einer schlechten Übereinstimmung, bei Werten von $0,4 < r_p < 0,7$ von einer mäßigen bis guten Übereinstimmung und bei Werten $r_p \geq 0,7$ von einer ausgezeichneten Übereinstimmung gesprochen. Zur Bewertung der Reproduzierbarkeit (intraindividuelle Reliabilität) wurde ebenfalls der Produkt-Moment-Korrelationskoeffizient nach Pearson (r_p) herangezogen.

4 Ergebnisse

4.1 Vergleich der Durchschnittsnoten verschiedener Bewertergruppen

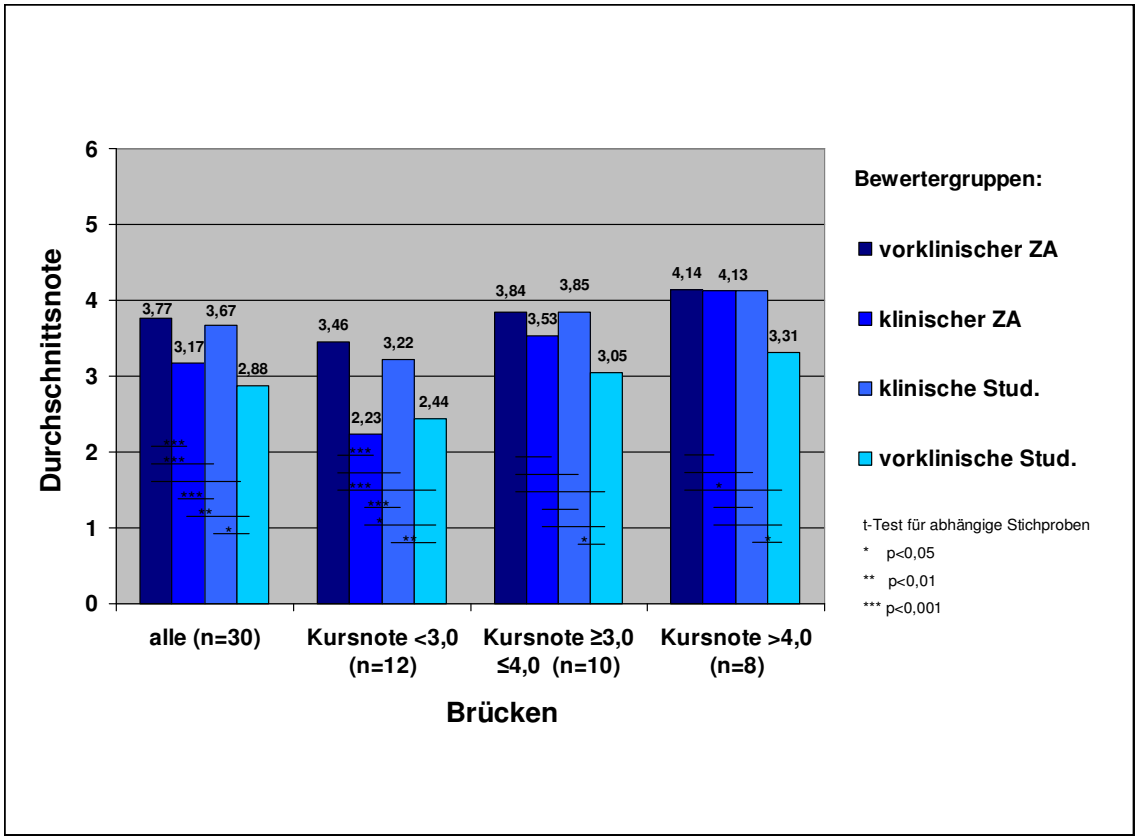

Abb. 4 *Durchschnittsnoten der Brücken.*

Abb. 4 zeigt die vergebenen Durchschnittsnoten der einzelnen Bewertergruppen für die gesamte Arbeit. Die vier Bewertergruppen sind farbig zu unterscheiden. Auf der y-Achse ist die Note abzulesen. Statistische Vergleiche der einzelnen Ergebnisse miteinander sind durch Verbindungslinien dargestellt. Statistisch signifikante Unterschiede sind durch Sternchen gekennzeichnet. Kein Sternchen bedeutet, dass ein vorhandener Unterschied statistisch nicht signifikant ist ($p \geq 0,05$). Die linke Balkengruppe zeigt die Durchschnittsnoten aller Arbeiten (n = 30). Die Durchschnittsnote der vorklinischen

Zahnärzte lag bei 3,77, die der klinischen Zahnärzte bei 3,17, die der klinischen Studenten bei 3,67 und die der vorklinischen Studenten bei 2,88. Die Unterschiede in den Durchschnittsnoten zwischen vorklinischen und klinischen Zahnärzten, zwischen vorklinischen Zahnärzten und klinischen Studenten und zwischen klinischen Zahnärzten und klinischen Studenten waren statistisch höchst signifikant.

Bei den drei nach rechts folgenden Balkengruppen sind die vergebenen Durchschnittsnoten in Abhängigkeit von der Kursnote der jeweiligen Arbeit dargestellt. Die Arbeiten der Kursnote < 3,0 (n = 12) wurden im Durchschnitt von den vorklinischen Zahnärzten mit 3,48, von den klinischen Zahnärzten mit 2,23, von den klinischen Studenten mit 3,22 und von den vorklinischen Studenten mit 2,44 bewertet. Statistisch höchst signifikant waren hier die Unterschiede zwischen vorklinischen Zahnärzten und klinischen Zahnärzten, vorklinischen Zahnärzten und vorklinischen Studenten und zwischen klinischen Zahnärzten und klinischen Studenten. Die im mittleren Leistungsdrittel angesiedelten Arbeiten (n = 10) wurden durchschnittlich von den vorklinischen Zahnärzten mit 3,84, von den klinisch tätigen Zahnärzten mit 3,53, von den klinischen Studenten mit 3,85 und von den vorklinischen Studenten mit 3,05 bewertet. Hier lag der einzige statistisch signifikante Unterschied zwischen den klinischen und den vorklinischen Studenten vor. Auf der rechten Seite der Grafik sind die Arbeiten mit der Kursnote > 4 dargestellt (n = 8). Von den vorklinischen Zahnärzten wurden diese im Durchschnitt mit der Note 4,14, von den klinischen Zahnärzten mit der Note 4,13, von den klinischen Studenten ebenfalls mit der Note 4,13 und von den vorklinischen Studenten mit der Note 3,31 bewertet. Statistisch signifikante Unterschiede wiesen die Vergleiche der vorklinischen Zahnärzte mit den vorklinischen Studenten und der klinischen mit den vorklinischen Studenten auf.

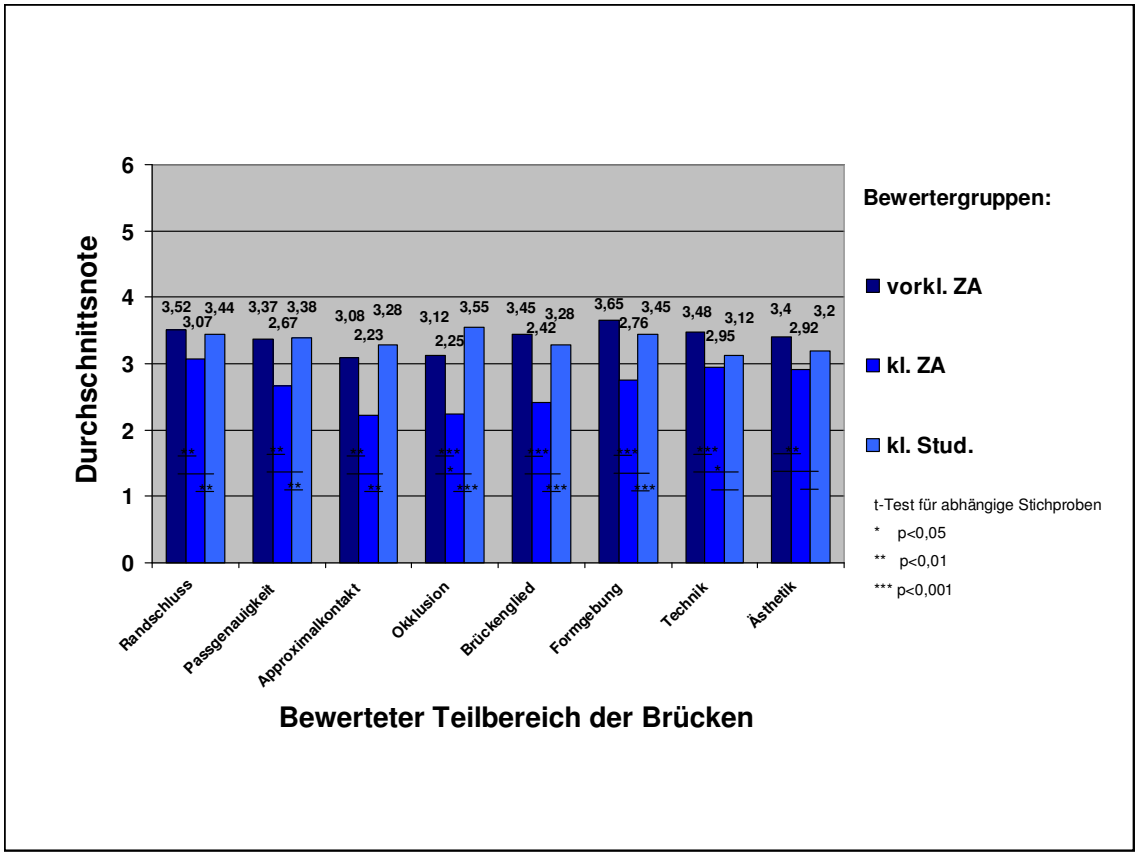

Abb. 5 *Durchschnittsnoten der Teilbereiche der Brücken.*

In Abb. 5 sind die Durchschnittsnoten nochmals nach den jeweils bewerteten Teilbereichen aufgeschlüsselt. Diese Teilbewertung wurde nur von den vorklinischen und klinischen Zahnärzten und den klinischen Studenten vorgenommen. Es handelt sich hier um die Durchschnittsnoten aller Arbeiten (n = 30). Für den Teilaspekt „Randschluss“ vergaben die vorklinischen Zahnärzte z. B. die Durchschnittsnote 3,52, die klinischen Zahnärzte die Note 3,07 und die klinischen Studenten die Note 3,44. Sowohl vorklinische Zahnärzte als auch klinische Studenten unterschieden sich hier von den klinischen Zahnärzten statistisch hochsignifikant. Noch deutlichere Unterschiede sieht man bei der Bewertung des Brückengliedes, bei welcher die vorklinischen Zahnärzte eine 3,45, die klinischen Zahnärzte eine 2,42 und die klinischen Studenten eine 3,28 gaben oder bei der Okklusions-Bewertung, bei der die vorklinischen Zahnärzte eine 3,12, die klinischen Zahnärzte eine 2,25 und die klinischen Studenten eine 3,55 vergaben. Bei

beiden Bewertungen unterschieden sich vorklinische Zahnärzte und klinische Studenten statistisch höchst signifikant von den klinischen Zahnärzten.

4.2 Notendifferenzen innerhalb der jeweiligen Bewertergruppe

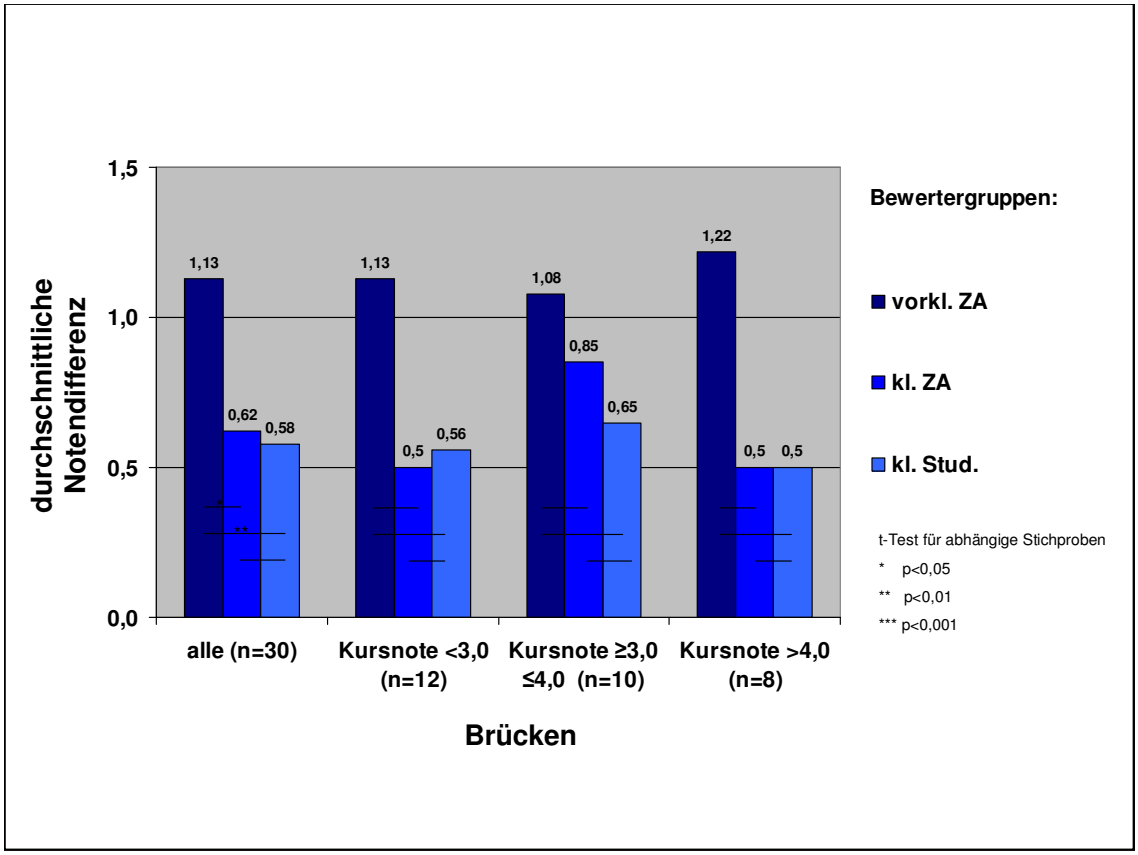

Abb. 6 *Notendifferenzen innerhalb der jeweiligen Bewertergruppe.*

Als Maß für die interindividuelle Übereinstimmung bei der Bewertung wurde innerhalb der verschiedenen Bewertergruppen jeweils die Differenz zwischen den durch die Bewerter vergebenen Noten gebildet. Die sich hieraus ergebenden durchschnittlichen Notendifferenzen innerhalb der verschiedenen Bewertergruppen sind in Abb. 6 dargestellt. Es sind sowohl die Notendifferenzen mit Blick auf die Durchschnittsnoten aller Arbeiten als auch die Notendifferenzen der Durchschnittsnoten differenziert nach Leistungsdritteln dargestellt.

Signifikante Unterschiede waren nur bei der Bewertung aller Brücken (n = 30) feststellbar. Hier unterschieden sich die vorklinischen Zahnärzte von den klinischen Zahnärzten statistisch signifikant und von den klinischen Studenten sogar statistisch hochsignifikant. Am stärksten waren die Abweichungen in der Notengebung stets innerhalb der Gruppe der vorklinischen Zahnärzte.

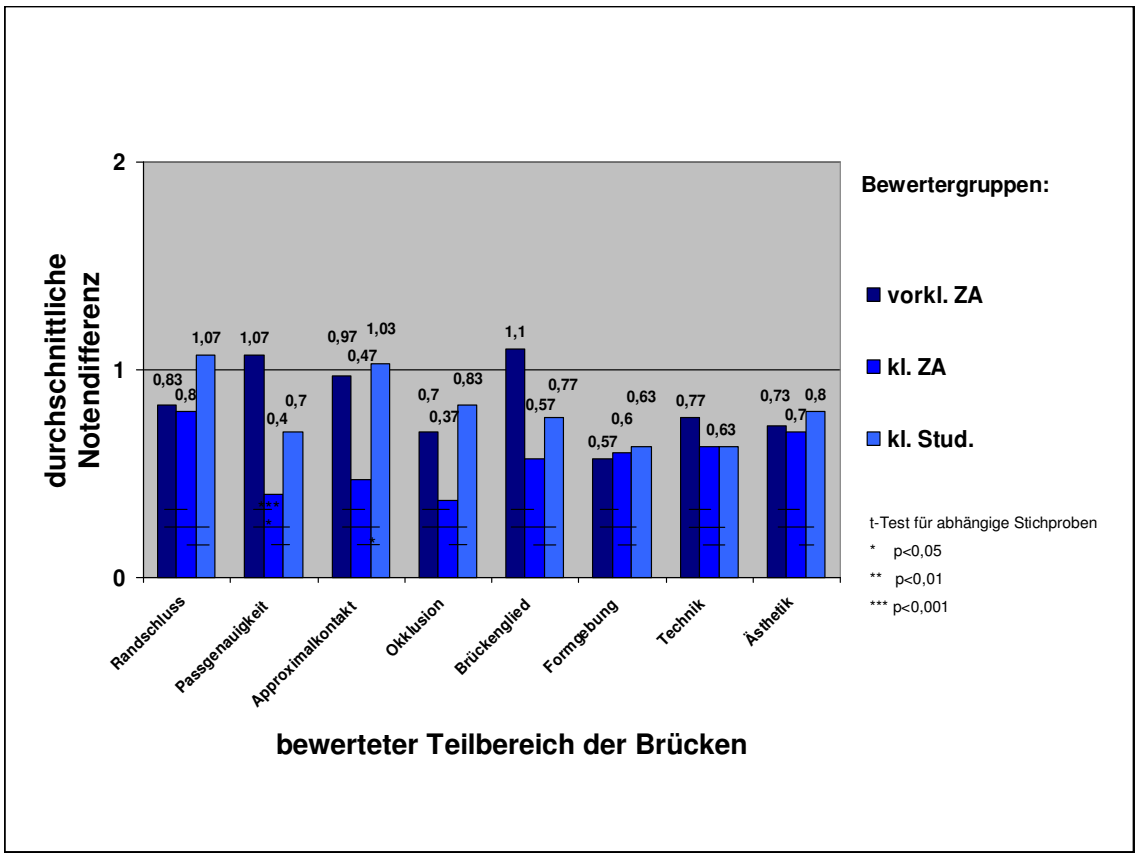

Abb. 7 *Notendifferenzen innerhalb der jeweiligen Bewertergruppe, Teilbereiche.*

Wie Abb. 6 zeigt auch Abb. 7 die durchschnittlichen Notendifferenzen innerhalb der jeweiligen Bewertergruppe, hier allerdings nach Teilbereichen differenziert. Es wiesen alle drei Bewertergruppen relativ ausgeprägte Differenzen auf.

Bei dem Qualitätskriterium „Randschluss“ differierten die Noten der klinischen Studenten um 1,07, bei der „Passgenauigkeit“ wiesen die vorklinischen Zahnärzte wiederum mit 1,07 die größte Notendifferenz auf. Mit Blick auf das Kriterium „Approximalkon-

takt“ zeigten sowohl die vorklinischen Zahnärzte mit 0,97 als auch die klinischen Studenten mit 1,03 deutliche Notendifferenzen in ihrer jeweiligen Gruppe. Die größte Notendifferenz von 1,1 trat bei der Bewertung des „Brückenglieds“ innerhalb der Gruppe der vorklinischen Zahnärzte auf (Abb. 4).

Die Unterschiede der durchschnittlichen Notendifferenz zwischen den Bewertergruppen waren größtenteils statistisch nicht signifikant. Der einzige statistisch signifikante Unterschied trat bei der Bewertung der Passgenauigkeit auf. Hier bestand in der Bewertungsabweichung ein statistisch höchstsignifikanter ($p < 0,001$) Unterschied zwischen vorklinischen und klinischen Zahnärzten, sowie ein statistisch signifikanter ($p < 0,05$) Unterschied zwischen vorklinischen Zahnärzten und klinischen Studenten. Für den Teilbereich „Approximalkontakt“ bestand ein statistisch signifikanter Unterschied in den Notendifferenzen zwischen klinischen Zahnärzten und klinischen Studenten.

4.3 Vergleich der studentischen Selbsteinschätzung mit dem Urteil anderer Bewerter

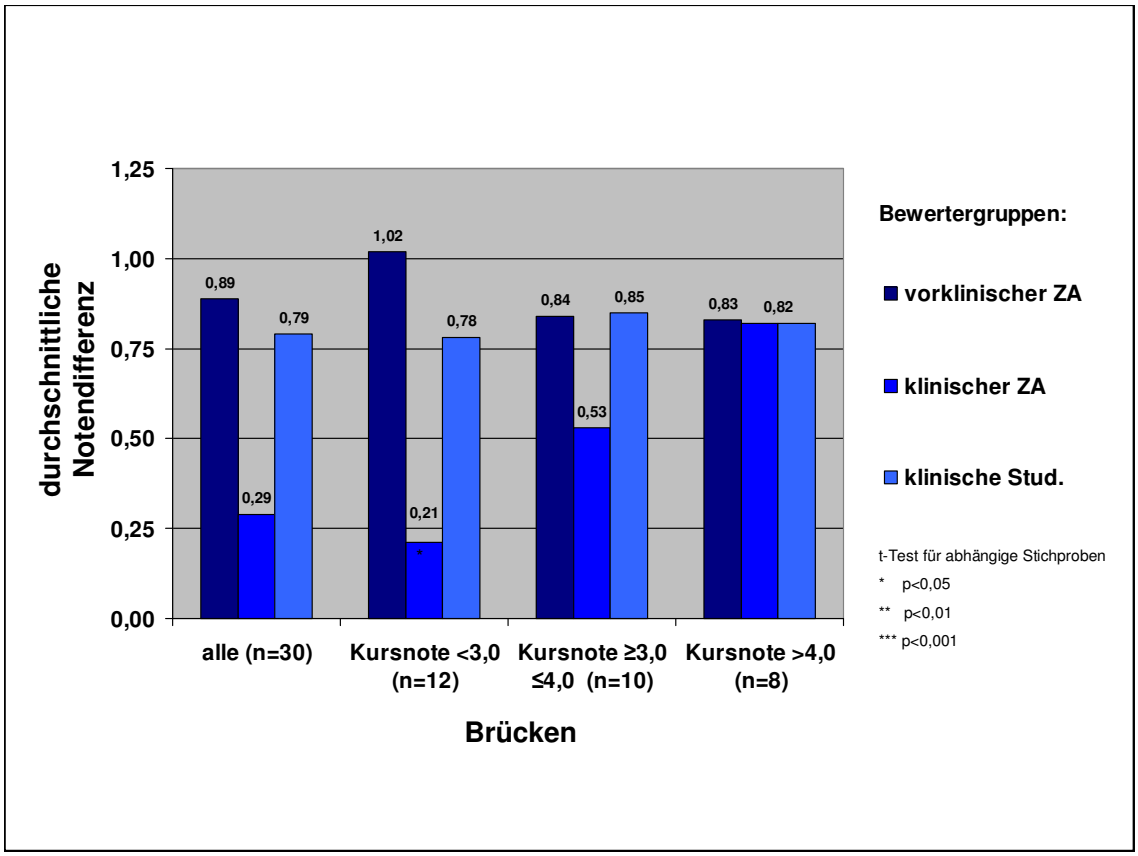

Abb. 8 *Vergleich der studentischen Selbsteinschätzung mit dem Urteil anderer Bewerter (Note der Selbsteinschätzung von Note der Fremdeinschätzung subtrahiert).*

Abb. 8 stellt den Unterschied zwischen der studentischen Selbsteinschätzung und dem Urteil der anderen Bewerter dar. Es sind die Unterschiede in der Durchschnittsnote sowohl für alle Arbeiten als auch für die nach Leistungsdritteln differenzierten Arbeiten abgebildet.

Grundsätzlich lagen die durchschnittlichen Notendifferenzen im Bereich > 0, was bedeutet, dass die Studierenden ihre eigene Arbeit durchweg besser bewerteten als die übrigen Bewerter. So wurden die Arbeiten der Kursnote < 3,0 (n = 12) von den vorklinischen Zahnärzten im Durchschnitt um die Note 1,02 strenger bewertet als von den vorklinischen Studenten selbst. Hier bestand der ausgeprägteste Unterschied. Die klini-

schen Zahnärzte bewerteten hier nur um die Note 0,21 strenger als die Studenten sich selbst einschätzten. Das stellte den geringsten Unterschied dar. Insgesamt lag die durchschnittliche Abweichung zwischen Selbst- und Fremdeinschätzung zwischen 0,21 und 1,02 (Abb. 8).

4.4 Interpersonelle Urteilskonkordanz innerhalb verschiedener Bewertergruppen

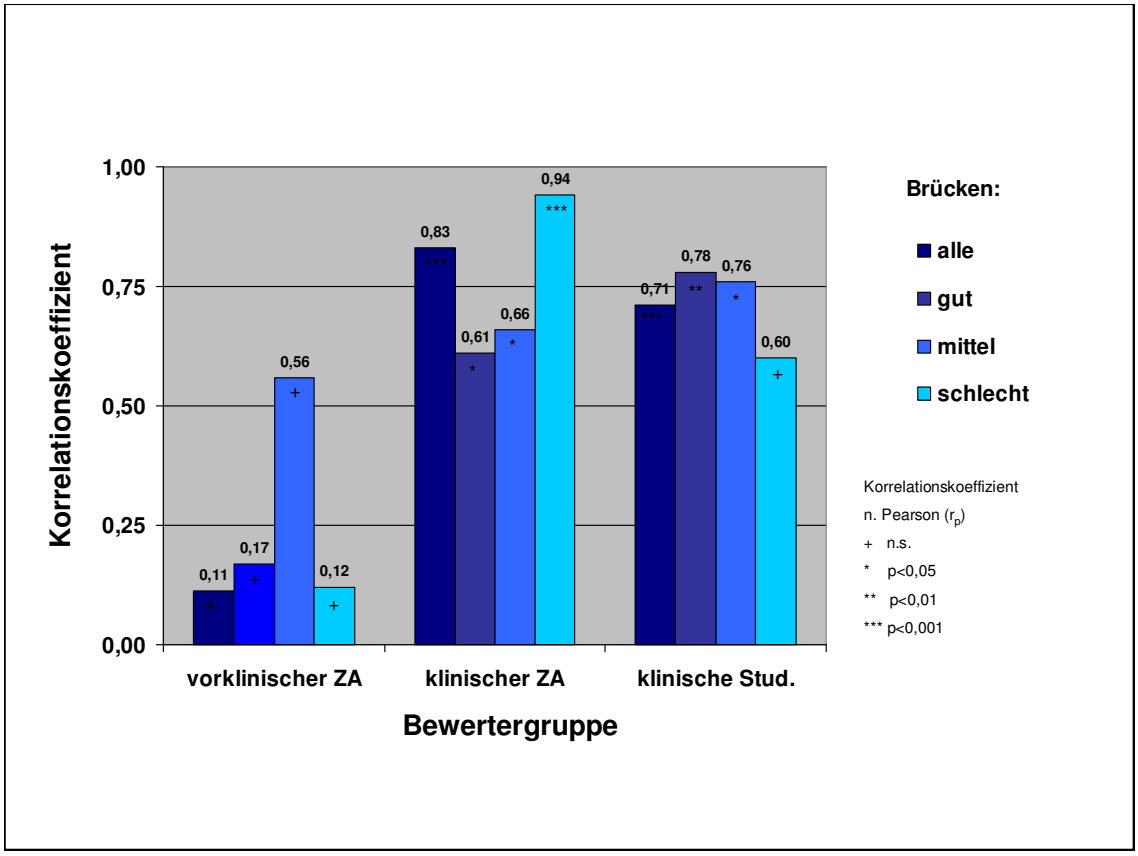

Abb. 9 *Interpersonelle Urteilskonkordanz innerhalb verschiedener Bewertergruppen.*

In Abb. 9 ist die interpersonelle Urteilskonkordanz der jeweiligen Bewerter in den verschiedenen Bewertergruppen dargestellt, wobei ein Korrelationskoeffizient von 1 eine 100 % - Übereinstimmung zwischen den zwei Bewertern der verschiedenen Gruppen

bedeutet. Wie man sieht, bestand innerhalb der Gruppe der vorklinischen Zahnärzte die geringste Korrelation bei der Bewertung, d. h. in dieser Gruppe lag die geringste interpersonelle Urteilskonkordanz vor. Am stärksten war in der Gruppe die Übereinstimmung noch bei der Bewertung der Brücken des mittleren Notendrittels mit einem Korrelationskoeffizienten von $r_p = 0{,}56$. Dieser Wert lag jedoch noch weit unter den Korrelationswerten der klinischen Zahnärzte und der klinischen Studenten.

Die größte interpersonelle Urteilskonkordanz lag bei den klinischen Zahnärzten bei der Bewertung der Arbeiten des unteren Notendrittels vor. Hier wurde ein Korrelationskoeffizient von 0,94, d. h. fast eine perfekte Übereinstimmung bei der Bewertung erreicht. Sowohl dieser, als auch die Korrelationskoeffizienten für die Bewertung aller Brücken (n = 30) durch die klinischen Zahnärzte und die klinischen Studenten waren statistisch höchst signifikant.

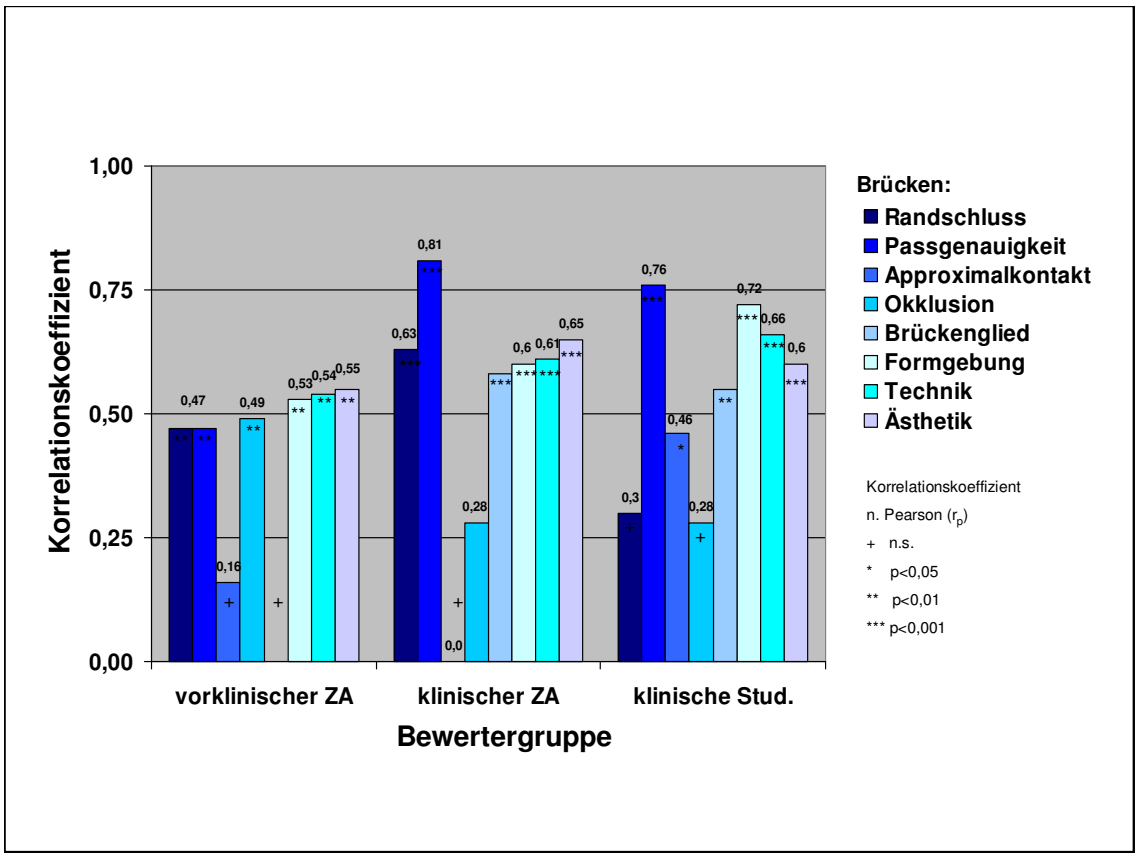

Abb. 10 *Interpersonelle Urteilskonkordanz innerhalb verschiedener Bewertungsgruppen in Bezug auf bewertete Teilaspekte der Brücken.*

Wie bereits in Abb. 9 beschrieben, zeigt auch Abb. 10 die Urteilskonkordanz der zwei Bewerter innerhalb der jeweiligen Bewertergruppen anhand des Korrelationskoeffizienten nach Pearson (r_p). In Abb. 9 wird die Urteilskonkordanz mit Blick auf die Gesamtnote aller Brücken und der Gesamtnoten der verschiedenen Leistungsdrittel dargestellt. In Abb. 10 hingegen werden keine Leistungsdrittel unterschieden. Es wird vielmehr die interpersonelle Reliabilität der Bewertung der acht verschiedenen Bewertungsaspekte dargestellt.

Die stärkste Übereinstimmung lag mit einem Korrelationskoeffizienten von 0,81 bei der Bewertung der Passgenauigkeit durch die klinischen Zahnärzte vor. Auch die klinischen Studenten erreichten bei der Passgenauigkeits-Bewertung mit 0,78 eine relativ hohe Korrelation. Die geringste Urteilskonkordanz lag bei den klinischen Zahnärzten mit einem Korrelationskoeffizienten von 0 bei der Bewertung des Approximalkontakts vor, d. h. hier bestand zwischen den beiden Bewertern grundsätzlich überhaupt keine Übereinstimmung. Korrelationskoeffizienten statistisch höchster Signifikanz wurden mehrfach in der Gruppe der klinischen Zahnärzte und in der Gruppe der klinischen Studenten erreicht. Ansonsten bestand durchweg eine hohe interpersonelle Urteilskonkordanz innerhalb der Gruppen der klinischen Zahnärzte und der klinischen Studenten.

4.5 Intrapersonelle Urteilskonkordanz

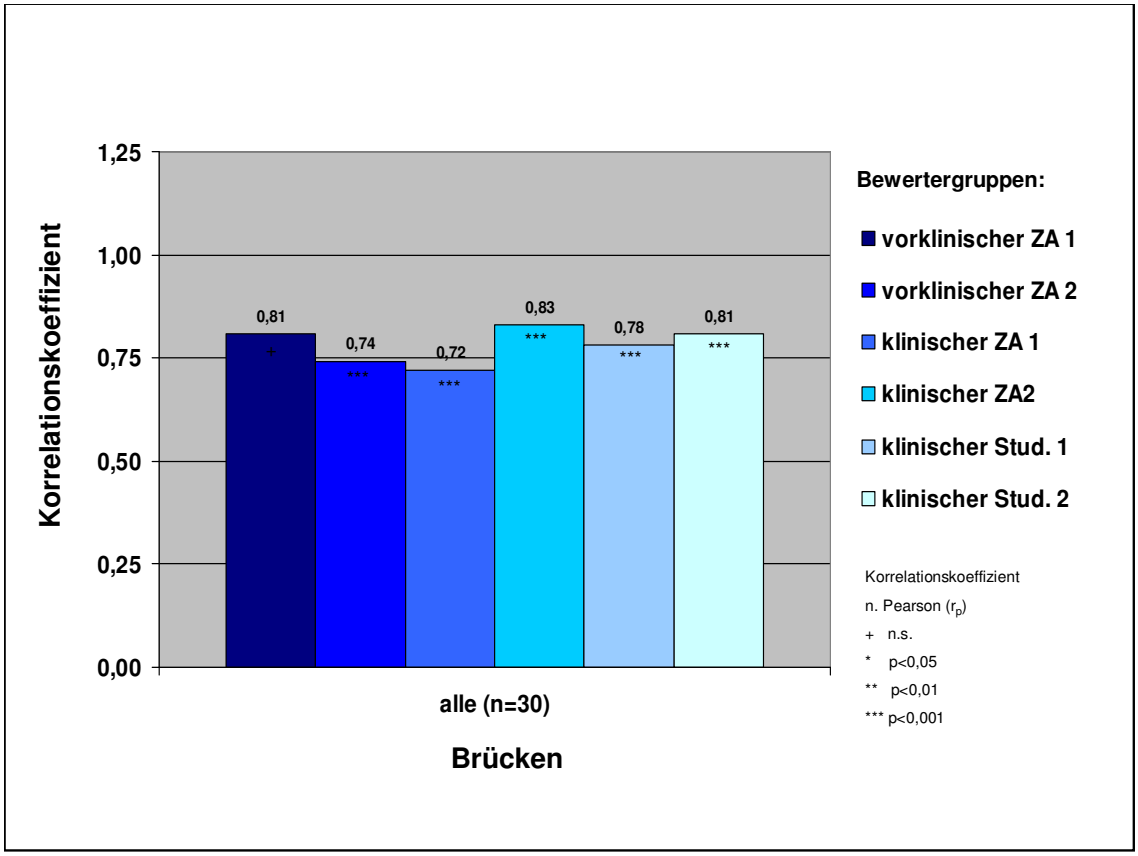

Abb.11 *Intrapersonelle Urteilskonkordanz bei wiederholter Bewertung durch dieselben Bewerter*

Abb. 11 zeigt die intrapersonelle Urteilskonkordanz, d. h. die Reproduzierbarkeit der Bewertung für alle sechs Bewerter.

Hierbei ist festzustellen, dass der klinische Zahnarzt 2 bei der Bewertung derselben Arbeiten am häufigsten gleich bewertete. Hier betrug der Korrelationskoeffizient 0,83. Grundsätzlich war allerdings bei allen Bewertern eine relativ hohe intraindividuelle Urteilskonkordanz ($r_p = 0,72 – 0,83$) zu beobachten ($p < 0,001$).

5 Diskussion

5.1 Eigene Ergebnisse

Die Frage nach der Reliabilität, d. h. der Zuverlässigkeit bzw. Objektivität und der Reproduzierbarkeit einer Diagnose, spielt in der klinischen Zahnmedizin eine wichtige Rolle. Ein typisches Beispiel aus dem klinischen Alltag ist die Befundung eines Patienten bei dessen Erstvorstellung. Nicht selten erscheint ein solcher Patient mit der Bitte um eine zweite Meinung und äußert sich dann sehr erstaunt über das, was ihm als Befund mitgeteilt wird. Füllungen, festsitzender oder herausnehmbarer Zahnersatz, Wurzelkanalbehandlungen oder auch parodontale Situationen werden plötzlich ganz anders beurteilt als durch den Vorbehandler. Die interindividuelle Reliabilität ist in einer solchen Situation also relativ gering. Diese Problematik ist weithin bekannt und führt häufig zur Verunsicherung des Patienten. Daher sollte bereits in der Ausbildung versucht werden, ein möglichst hohes Maß an Reliabilität zu erreichen. In diesem Zusammenhang ist es sinnvoll, die angehenden Zahnärzte zur kritischen und reproduzierbaren Eigenbewertung anzuleiten. Hierdurch wird für sie die Bewertung ihrer studentischen Leistung verständlicher und der Anspruch an die Qualität und deren Sicherung gefördert.

Die vorliegende Studie zeigt, dass bei dieser Eigenbewertung die Studierenden ihre Arbeiten in der Regel deutlich besser benoten als dies die übrigen Bewerter tun. Die studentische Selbsteinschätzung für die Bewertung der jeweiligen Brücke war im Durchschnitt um fast eine Note besser als die Noten der anderen. Erstaunlicherweise vergaben die klinischen Zahnärzte ebenfalls bessere Noten als die vorklinischen Zahnärzte und die klinischen Studenten. Die Differenz der Noten von klinischen Studenten und klinischen bzw. vorklinischen Zahnärzten war bei der Bewertung von Arbeiten des oberen Leistungsdrittels am ausgeprägtesten. Bei den Arbeiten des unteren Leistungsdrittels, welche unter der Bestehensgrenze lagen, war sie bei diesen drei Bewertergruppen nicht mehr vorhanden. Daraus lässt sich für die Lehre und den klinischen Alltag schlussfolgern, dass unbrauchbare Arbeiten in der Regel nicht zum Bestehen des Kurses führen

und dass qualitativ nicht ausreichende Arbeiten im Patientenmund auch als solche bewertet werden. Bei der Bewertung der Arbeiten des oberen Leistungsdrittels fällt auf, dass diese durch die vorklinischen Zahnärzte durchweg deutlich kritischer bewertet wurden (d. h. mit schlechteren Noten) als von den klinischen Zahnärzten. Dies ist wahrscheinlich auf ein höheres Qualitätsniveau in den vorklinischen Phantomkursen im Vergleich zu klinischen Patientenkursen zurückzuführen. Die sich hierin widerspiegelnden hohen Anforderungen - etwa an Passgenauigkeit, Randschluss, Approximalkontakt etc. - sind in den vorklinischen Kursen durchaus gerechtfertigt, da die Studierenden eine optimale Versorgung erlernen sollen und keine erschwerenden, patientenspezifischen Faktoren hinzukommen. Die klinischen Zahnärzte weisen den größten Erfahrungsschatz auf. Sie haben häufiger suboptimale, jedoch noch funktionstüchtige prothetische Arbeiten im Patientenmund gesehen und bewerten somit nahezu optimal gestaltete Brücken besonders positiv. Sie kennen Situationen, in denen Kompromisse bezüglich der Qualität eingegangen werden müssen, da die Patientensituation anderes nicht zulässt.

Betrachten wir nun die Untersuchung bezüglich der interindividuellen Reliabilität in den jeweiligen Bewertergruppen, so unterschieden sich die vorklinischen Zahnärzte mit Notendifferenzen zwischen 1,08 und 1,22 am deutlichsten. Diese Differenz ist bei der Bewertung aller Leistungsbereiche zu beobachten. Zwischen den beiden klinischen Studenten bestand insgesamt die geringste Notenabweichung. Als Grund hierfür ist der fast identische Wissensstand beider Bewerter zu vermuten. Sie haben zum Untersuchungszeitpunkt alle Kurse gemeinsam absolviert und bei Gruppenarbeiten häufig zusammen gearbeitet. Es wies also keiner der beiden einen gravierenden Wissensvorsprung zum anderen auf.

Statistisch signifikante Unterschiede zwischen den Notendifferenzen der jeweiligen Gruppen waren nur bei der Bewertung aller Brücken (n = 30) zu verzeichnen. Hier unterscheidet sich die Differenz der vorklinischen Zahnärzte signifikant von der der klinischen Zahnärzte und hochsignifikant von der der klinischen Studenten. Betrachtet man nun die Notendifferenzen für die bewerteten Teilbereiche der Brücken, so wiesen auch die klinischen Studenten für die Bereiche „Randschluss“ und „Approximalkontakt“ Differenzen auf, welche die Größe einer Notenbandbreite übersteigen. Nur beim „Approximalkontakt“ unterschied sich diese Differenz signifikant von der Differenz der

klinischen Zahnärzte. Für den Teilbereich „Passgenauigkeit“, der von den vorklinischen Zahnärzten mit einer Notendifferenz von 1,07 bewertet wurde, lagen ein höchst signifikanter Unterschied der Differenzen zu den klinischen Zahnärzten und ein signifikanter Unterschied zu den klinischen Studenten vor.

Sieht man nun die Notendifferenz innerhalb einer Bewertergruppe als Maß für die interpersonelle Reliabilität von Zahnärzten ungefähr gleichen Wissensstands, so wiesen die vorklinischen Zahnärzte die geringste Reliabilität auf. Ein Versuch der Erklärung hierfür kann von verschiedenen Standpunkten aus beginnen. Die demnach höhere Reliabilität bei den klinischen Assistenten wäre durch ein besonders hohes Maß an klinischer Erfahrung, welches erst nach einer längeren Tätigkeitsdauer erreicht wird, zu erklären. Ab diesem Wissensstand findet ein Wissenszuwachs langsamer und eher in Spezialbereichen statt, der aber zur Beurteilung der untersuchten Arbeiten nicht bedeutsam ist. Somit befinden sich beide klinischen Zahnärzte auf etwa demselben Erfahrungsstand und stellen ähnliche Qualitätsanforderungen. Die geringe Notendifferenz der klinischen Studenten wurde bereits oben erklärt. Andererseits kann man die geringen Notendifferenzen auch als erwartet ansehen und die größeren Differenzen der vorklinischen Assistenten mit größeren personellen Unterschieden erklären. Besonders zu Beginn der beruflichen Laufbahn machen sich ein Erfahrungsvorsprung etwa durch häufigeres Behandeln oder Eigenstudium bemerkbar. Einer der beiden vorklinischen Zahnärzte könnte sich einen solchen Erfahrungsvorsprung erarbeitet haben. Eine mögliche Erklärung wären auch persönliche Unterschiede in den Anforderungen an die Qualität der eigenen Arbeit und der anderer.

Bewertet man die interpersonelle Urteilskonkordanz innerhalb der verschiedenen Bewertergruppen statt über die Notendifferenz mittels eines Korrelationskoeffizienten, zeigt sich ein ähnliches Ergebnis innerhalb der jeweiligen Bewertergruppe. So erreichten die vorklinischen Zahnärzte im Durchschnitt nur eine sehr geringe Korrelation von $r_p = 0{,}11$ ($p > 0{,}05$). Eine sehr hohe Korrelation mit $r_p = 0{,}83$ ($p < 0{,}001$) zeigten dagegen die klinischen Zahnärzte. Die klinischen Studenten wiesen ebenfalls durchweg einen hohen Korrelationskoeffizienten ($r_p = 0{,}6 - 0{,}78$, $p < 0{,}05$) auf, d. h. auch hier kann man von einer ausgeprägten Urteilskonkordanz sprechen.

Die interpersonelle Urteilskonkordanz bei der Bewertung der Teilbereiche zeigt, dass sowohl vorklinische als auch klinische Zahnärzte bei der Bewertung des Approximalkontaktes und klinische Zahnärzte und klinische Studenten bei Okklusionsbewertung nicht signifikant korrelierten. Diese Bewertungskriterien scheinen individuellen Spielraum zu lassen und bedürfen genauerer Definition. Ein Gegenbeispiel hierzu ist die Bewertung der Passgenauigkeit. Hier wurden hoch- und höchstsignifikante Korrelationskoeffizienten erreicht. Die Beurteilung der Passgenauigkeit gleicht einer ja / nein-Entscheidung. Hier bleibt wenig Ermessensspielraum. Obwohl bei Formgebung, Technik und Ästhetik der Ermessensspielraum größer erscheint, werden hier ebenso gute Übereinstimmungen erreicht.
Bei der intrapersonellen Urteilskonkordanz, d. h. der Reproduzierbarkeit bei Mehrfachbewertung durch dieselbe Person, zeigten alle Bewerter sehr hohe Korrelationskoeffizienten, die bis auf den vorklinischen Zahnarzt 1 statistisch höchst signifikant waren. Dieses Ergebnis lässt die Vermutung zu, dass die intrapersonelle Reliabilität nicht vom Erfahrungsgrad des Bewerters abhängt.

5.2 Vergleich mit der Literatur

Betrachtet man abschließend die in vorliegender Studie erzielten Ergebnisse vor dem Hintergrund bisheriger Untersuchungen zur Reliabilität bei der Bewertung studentischer Arbeiten, ist Folgendes festzustellen:
Die Kalibrierung verschiedener Bewerter verbessert deutlich deren interindividuelle Reliabilität. Dies spiegelt sich in den guten Übereinstimmungen der klinischen Studenten wider. Die Literatur bietet sowohl positive als auch negative Ergebnisse für die Reliabilitätssteigerung durch Bewerterkalibrierung [1, 7, 14, 23], jedoch beschrieben beispielsweise *Abou-Rass* [1] und *Natkin und Guild* [23] eine Reliabilitätssteigerung durch Kalibrierungsversuche wie Trainings- bzw. Diskussionstreffen. Die lange Zusammenarbeit der klinischen Studenten ist durchaus mit einem solchen Training vergleichbar und spricht für dessen Erfolg.

Vor dem Hintergrund der Reliabilitätssteigerung in Abhängigkeit der Berufserfahrung bestätigen die Resultate der vorliegenden Untersuchung die Ergebnisse der Literatur [4, 23]. Die erfahrenen klinischen Assistenten erreichten besonders positive Werte bei der interindividuellen Reliabilität (Objektivität). Ähnliches wurde von *Natkin und Guild* [23] und von *Dhuru et al.* [4] für Bewerter mit hohem Erfahrungsniveau beschrieben.
Vergleicht man nun die errechneten Korrelationskoeffizienten mit denen der Literatur, so fallen auf beiden Seiten ausgeprägte Schwankungen der Werte auf. Beispielsweise errechneten *Lilley et al.* [19] Korrelationskoeffizienten von $r_p = 0{,}11 – 0{,}72$ für die Objektivität und von $r_p = 0{,}51$ - $0{,}63$ für die Reproduzierbarkeit. Vorliegende Studie ergab Korrelationskoeffizienten von $r_p = 0{,}11 – 0{,}56$ für die Objektivität der vorklinischen Zahnärzte und von $r_p = 0{,}61 – 0{,}94$ für die Objektivität der klinischen Zahnärzte (siehe Abb. 6). Bei der Reproduzierbarkeit der Bewertung erreichten alle positive Korrelationskoeffizienten von $r_p = 0{,}72 – 0{,}83$ (siehe Abb. 8). Folglich sind die in aktueller Untersuchung erreichten Korrelationskoeffizienten, insbesondere die der erfahrenen bzw. gut kalibrierten Bewerter als äußerst positiv im Vergleich zu den Werten der Literatur anzusehen [2, 4, 6, 7, 8, 9, 13, 14, 19, 22, 28].
Im Hinblick auf den Einfluss der Art des zu bewertenden Punktes fällt insbesondere die Bewertung der Passgenauigkeit der Brücken durch die klinischen Zahnärzte bzw. klinischen Studenten mit Korrelationskoeffizienten von $r_p = 0{,}81$ und $r_p = 0{,}76$ auf (siehe Abb. 7). Diese äußerst positiven Werte lassen sich durch die Einfachheit des Bewertungspunktes im Sinne einer ja-nein-Entscheidung erklären (Schaukeln vorhanden oder nicht). Andere Autoren nutzten dieses Phänomen beim Versuch, ein simples Bewertungssystem zu entwickeln [13, 14].

5.3 Schlussfolgerungen

Es ergeben sich Konsequenzen in zweierlei Hinsicht. Zum einen soll eine möglichst objektive und reproduzierbare Bewertung in der Ausbildung ermöglicht werden, zum anderen soll die Qualität der zahnärztlichen Arbeit von gleich bleibend hohem Niveau sein.

In der vorliegenden Studie, wurden nur innerhalb der Gruppe der klinischen Zahnärzte und der klinischen Studenten akzeptable Korrelationskoeffizienten > 0,7 erreicht, wogegen in der Gruppe der vorklinischen Kursassistenten nur eine geringgradige Übereinstimmung (Objektivität) vorhanden war.

Dies zeigt, dass selbst relativ unerfahrene Bewerter, wie die klinischen Studenten, hervorragende objektive Bewertungen abgeben können, wenn sie untereinander gut kalibriert sind. Diese Bewerterkalibrierung mit genauer Definition von Kriterien und dem Training der Anwendung dieser Kriterien ist ein sinnvolles Hilfsmittel, um angemessene Urteile zu finden. So wären selbst unerfahrene Bewerter in der Lage, objektive Bewertungen abzugeben, was die positiven Ergebnisse der klinischen Studenten belegen. Besonders in der Ausbildung könnten diese Maßnahmen hilfreich sein. Zusätzlich würde die Selbsteinschätzung erleichtert. Diese erwies sich in der vorliegenden Studie als äußerst unrealistisch. Die vorklinischen Studenten bewerteten sich selbst mehr als eine Note besser als die anderen Bewerter und hoben alle Noten über die Bestehensgrenze.

Neben der Festlegung der leistungsspezifischen Anforderungen erscheint auch die Bewertung durch eine zweite Person sinnvoll. Besonders bei Arbeiten unterhalb oder im Grenzbereich der Bestehensgrenze ist ein Vergleich der Note mit der Bewertung durch den Kursleiter anzustreben, um somit die größere Berufserfahrung mit in die Bewertung einfließen lassen zu können. Die hohen Korrelationskoeffizienten der klinischen Zahnärzte bestätigen den Einfluss des Faktors „Berufserfahrung“, was ein zweites wesentliches Ergebnis der Studie darstellt.

Weiterhin wichtig ist die Tatsache, dass das Eigenurteil aller Bewerter durchaus reproduzierbar war und somit in diesem Bereich in der Ausbildung und in der täglichen Behandlung nur geringe Defizite vorliegen. Ebenfalls positiv ist die hohe Übereinstimmung bei der Bewertung der schlechten Brücken (Note > 4). Dies zeigt, dass die Mindestanforderungen an die Qualität einer Restauration auf gleichem Niveau liegen.

Für die Zukunft sollte versucht werden, durch einfache Hilfsmittel wie strukturierte Bewertungsbögen mit definierten Notenvorgaben und Vereinfachungen der zu bewertenden Punkte im Sinne von ja-nein-Entscheidungen die Objektivität und die Reproduzierbarkeit weiter zu steigern.

6 Zusammenfassung

In der Medizin ergibt sich grundsätzlich die Notwendigkeit der Leistungsbewertung beziehungsweise der Beurteilung klinischer Situationen oder der Ergebnisqualität.
Klassische Gütekriterien für die Reliabilität einer Bewertung sind die Objektivität (interindividuelle Reliabilität) und die Reproduzierbarkeit (intraindividuelle Reliabilität). Das Kriterium Objektivität (interindividuelle Reliabilität) gibt an, wieweit das Ergebnis eines Tests abhängig davon ist, wer die Untersuchung durchgeführt hat. Bei der Reproduzierbarkeit (interindividuelle Reliabilität) geht es um die Wiederholbarkeit und damit Zuverlässigkeit der Datenerhebung in Anlehnung an das Messkonzept in der Physik, wobei man davon ausgeht, dass die zu messenden Merkmale über die Zeit stabil sind.
Was die Beurteilung vorklinischer studentischer Arbeiten in der Zahnmedizin betrifft, ist in erster Linie die Reliabilität (Zuverlässigkeit) der Bewertung/Benotung von Bedeutung.
Ziel der vorliegenden Untersuchung war es, das Ausmaß der in der Person des Bewertenden begründeten inter- und intrapersonellen Variabilität bei der Benotung praktischer vorklinischer Studentenarbeiten am Phantomkopf (hier: Kunststoffverblendbrücke 24 - 26) festzustellen. In diesem Zusammenhang sollte untersucht werden, ob ein statistisch signifikanter Unterschied zwischen verschiedenen Bewertergruppen mit unterschiedlicher Berufserfahrung (vorklinische Studenten, klinische Studenten, vorklinische Zahnärzte, klinische Zahnärzte) besteht und inwieweit die Selbst- bzw. Fremdeinschätzung von Studierenden des vorklinischen Phantomkurses verlässlich ist.
Die Bewertung fand durch vier Gruppen unterschiedlichen Erfahrungsniveaus statt. Zwei Assistenten der klinischen Studentenkurse, zwei Assistenten der vorklinischen Studentenkurse und zwei Studenten des klinischen Studienabschnittes bewerteten 30 Kunststoffverblendbrücken des Phantomkurses II. Zusätzlich gaben die vorklinischen Studenten ihre Selbsteinschätzung ab. Die anonymisierten Arbeiten unterschiedlichen Leistungsniveaus wurden unter standardisierten Bedingungen bewertet. Es wurden sowohl die Gesamtleistung als auch acht Einzelpunkte bewertet. Als Grundlage dieser Bewertung diente, wie in der Zahnmedizin üblich, eine Checkliste (Aufzählung von Merkmalen, die einen Gegenstand umfassend beschreiben), die alle zu untersuchenden Punkte enthält, jedoch keine Kriterien vorgibt. Zwei Bewertungsdurchgänge mit einem

zeitlichen Abstand von mehr als 24 Stunden dienten zur Ermittlung der intraindividuellen Reliabilität. Die interindividuelle Reliabilität lässt sich aus dem Vergleich der einzelnen Gruppen bzw. der Einzelbewertungen errechnen.

Die Ergebnisse zeigten deutliche Unterschiede in der Notenvergabe. Die Eigenbewertung durch die vorklinischen Studenten und die Bewertung durch die klinischen Zahnärzte waren deutlich milder als die Bewertung durch klinische Studenten und vorklinische Zahnärzte. Diese Tendenz zeigte sich sowohl bei der Gesamtbewertung der Arbeit als auch bei der Bewertung einzelner Teilbereiche. Die Differenz der Durchschnittsnoten war jedoch bei den Arbeiten des mittleren und des schlechten Kursdrittels deutlich geringer als bei den Arbeiten des oberen Kursdrittels. Allerdings lag die Eigenbewertung der vorklinischen Studenten immer deutlich oberhalb der Bestehensgrenze.

In Hinsicht auf die Reproduzierbarkeit des eigenen Urteils unterschieden sich fünf der sechs Bewerter kaum. Sie wiesen alle eine statistisch höchst signifikante intraindividuelle Reliabilität auf. Einzig die intraindividuelle Reliabilität des vorklinischen Zahnarztes 1 war statistisch nicht signifikant.

Lenkt man den Blick auf die interindividuelle Urteilskonkordanz (Objektivität) der Bewertergruppen, so stachen die vorklinischen Zahnärzte mit einer nicht signifikanten Urteilskonkordanz und einer deutlichen Notendifferenz innerhalb ihrer Gruppe hervor.

Die positiven Ergebnisse der klinischen Studenten für die Objektivität der Bewertung zeigten, dass die Kalibrierung der Bewerter eine wichtige Rolle spielt. Durch ihr gemeinsames Studium waren sie sehr gut kalibriert. In der Literatur wurden bereits Versuche zur Reliabilitätssteigerung durch bessere Bewerterkalibrierung beschrieben [1, 7, 14, 23].

Des Weiteren belegen die Ergebnisse der klinischen im Vergleich zu denen der vorklinischen Assistenten den positiven Einfluss der großen Berufserfahrung auf die Objektivität der Beurteilung. Ein Vergleich der erreichten Korrelationskoeffizienten von $r_p = 0,72 - 0,83$ für die Reproduzierbarkeit und von $r_p = 0,62 – 0,94$ für die Objektivität der Bewertungen der klinischen Assistenten mit den in der Literatur beschriebenen Werten bestätigt die hohe Reliabilität ihrer Bewertung [2, 4, 6, 7, 8, 9, 13, 14, 19, 22, 28].

Weitere Hilfsmittel zur objektiven und reproduzierbaren Bewertung sollten, bereits im Studium, als sinnvoll angesehen und eingesetzt werden. So könnte einerseits die Eigen-

bewertung und damit das Verständnis der eigenen Leistung erleichtert und andererseits die Qualität der zahnärztlichen Arbeit zugesichert werden.

7 Literaturverzeichnis

1. Abou-Rass, MA (1973)
   Clinical Evaluation Instrument in Endodontics.
   J Dent Educ 37: 22-36.

2. Bedi R, Lo E, King NM, Chan T (1987)
   The effect of pictorial criteria upon the reliability of assessments of cavity preparations.
   J Dent 15: 222-224.

3. Billy EJ, Brandau HE., Fitzgerald M, Kolling JN, Lorey RL, Sloan KM (Hrsg.)
   Clinical Foundation II # 621 (Kursskript). The University of Michigan Dental Publications, Ann Arbour, MI 1997, vii-viii.

4. Dhuru VB, Rypel TS, Johnston WM (1982)
   Criterion-oriented grading system for preclinical operative dentistry laboratory course.
   J Dent Educ 42: 528-531.

5. Edwards WS, Morse PK, Mitchell RJ (1982)
   A Practical Evaluation System for Preclinical Restorative Dentistry
   J Dent Educ 46: 693-696.

6. Feil PH (1982)
   An analysis of the reliability of a laboratory evaluation system.
   J Dent Educ 46: 489-494.

7. Fuller JL (1972)
   The effects of training and criterion models on interjudge reliability.
   J Dent Educ 36: 19-22.

8. Gaines WG, Bruggers H, Rasmussen RH (1974)
Reliability of ratings in preclinical fixed prosthodontics: effect of objective scaling.
J Dent Educ 38: 672-675.

9. Goepferd SJ, Kerber PE (1980)
A comparison of two methods for evaluating primary class II cavity preparations.
J Dent Educ 44: 537-542.

10. Guild RE (1966)
An experiment in modified programmed self-instruction.
J Dent Educ 30: 181-9.

11. Guild RE (1966)
Questionnaire studies at three schools of dentistry.
J Dent Educ 30: 344-353.

12. Heffer P, Holloway PJ, Rose JS, Swallow JN (1965)
An Investigation into Dental Undergraduate Examining Techniques.
Br Dent J 118: 334-338.

13. Hinkelmann KW, Long NK (1973)
Method for decreasing subjective evaluation in preclinical restorative dentistry
J Dent Educ 37: 13-18.

14. Houpt MI, Kress G (1973)
Accuracy of measurement of clinical performance in dentistry.
J Dent Educ 37: 34-46.

15. Ingenkamp K (1995) Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte. 9. Aufl. Beltz, Weinheim.

16. Jenkins SM, Dummer PM, Gilmour AS, Edmunds DH, Hicks R, Ash P (1998):
Evaluating undergraduate preclinical operative skill; use of a glance and grade marking system.
J Dent 26: 679-684.

17. Kellersmann T. 2007 Zur Reliabilität der Beurteilung vorklinischer Phantomarbeiten bei Einsatz eines strukturierten Bewertungsbogens [Dissertation]. Westfälische Wilhelms-Universität, Münster

18. Killip DE (1965)
Role of discovery in learning manual skills.
J Dent Educ 29: 63-70.

19. Lilley JD, ten Bruggen Cate HJ, Holloway PJ, Holt JK, Start KB (1968)
Reliability of practical tests in operative dentistry.
Br Dent J 125: 194-197.

20. Mackenzie RS (1973)
Defining clinical competence in terms of quality, quantity, and need for performance criteria.
J Dent Educ 37: 37-44.

21. McDonald GT, Larsen HD (1988)
A system for developing and evaluating the clinical judgment of dental students.
J Prosthet Dent 53: 265-266.

22. Meetz HK, Bebeau MJ, Thoma SJ (1988)
The validity and reliability of a clinical performance rating scale.
J Dent Educ 52: 290-297.

23. Natkin E, Guild RE (1967)
Evaluation of preclinical laboratory performance: a systematic study.
J Dent Educ 31: 152-161.

24. Nedelsky L (1965)
Science Teaching and Testing
Harcourt, Brace, and World, Inc., Chicago, p.160.

25. O´Connor P, Lorey RE (1978)
Improving interrater agreement in evaluation in dentistry by the use of comparison stimuli
J Dent Educ 42: 174-179.

26. Robertello FJ, Pink FE (1997)
The effect of a training program on the reliability of examiners evaluating amalgam restorations.
Oper Dent 22: 57-65.

27. Ryge G, Snyder M (1973)
Evaluating the Clinical Quality of Restorations.
J Am Dent Assoc 87: 369-377.

28. Sachs L, Hedderich J (2003)
Angewandte Statistik. Anwendung statistischer Methoden, Springer, Berlin, 11. Aufl.

29. Salvendy G, Joost MG, Cunningham PR, Ferguson GW, Wilko RA, Dees RW (1976)
Improving evaluation of amalgam restorations.
J Dent Educ 40: 368-369.

30. Türp JC, Gerds Th, Schneider U (2002)
Variabilität bei der Benotung studentischer Arbeiten im vorklinischen Phantomkurs
Dtsch Zahnärztl Z 57: 526-531.

31. Vanek HG (1964)
Objective Evaluation of Dental Student Technic Products
J Dent Educ 33: 140-144.

32. Vanek HG, Chen MK, Podshadley DW (1967)
Evaluation of a Self-instructional Method Used in Preclinical Operative Dentistry
J Dent Educ 31: 34-43.

33. Vann WF, Machen JB, Hounshell PB (1983)
Effects of criteria and checklists on reliability in preclinical evaluation.
J Dent Educ 47: 671-675.

Lebenslauf

Name	Christian Philipp Schiffler	
Geburtsort	Duisburg	
Geburtsdatum	13.04.1979	
Familienstand	ledig	
Eltern	Wolfgang Schiffler	Jurist
	Monika Schiffler	MTA
Schulausbildung	1985 – 1989	Städtische Gemeinschaftsgrundschule an der van-Gogh-Straße, Duisburg-Trompet
	1989 – 1998	Gymnasium in den Filder Benden, Moers Abschluss mit dem Abitur
Studium	04 / 1999	Beginn des Studiums der Zahnmedizin an der Westfälischen Wilhelms-Universität Münster
	2000	naturwissenschaftliche Vorprüfung
	2001	zahnärztliche Vorprüfung
	2004	Staatsexamen, Approbation am 09.07.2004
	10 / 2006	Beginn des IMC-Masterstudiengangs für orale Medizin in Implantologie, Abschluss voraussichtlich 10 / 2007
Beruf	Seit September 2004 wissenschaftlicher Mitarbeiter der Poliklinik für Zahnerhaltung des Zentrums für Zahn- Mund- und Kieferheilkunde des Universitätsklinikums Münster	

Münster, den 08.06.2007