Aus dem Universitätsklinikum Münster
Klinik und Poliklinik für Zahnärztliche Prothetik
- Direktor: Univ.-Prof. Dr. med. Dr. med. dent. F. Bollmann -

Zur Reliabilität der Beurteilung vorklinischer Phantomarbeiten bei Einsatz eines strukturierten Bewertungsbogens

INAUGURAL-DISSERTATION

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

vorgelegt von
Kellersmann, Christian Tim
aus Hagen

2008

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:	Univ.-Prof. Dr. E. Schäfer
Tag der mündlichen Prüfung:	07. Februar 2008

Aus der Poliklinik für Zahnersatz des Zentrums für Zahn-, Mund- und Kieferheilkunde
der Westfälischen Wilhelms-Universität Münster
Direktor: Univ.-Prof. Dr. Dr. F. Bollmann
Referent: Univ.-Prof. Dr. P. Scheutzel
Korreferent: Univ.-Prof. Dr. E. Schäfer

Zusammenfassung

Zur Reliabilität der Beurteilung vorklinischer Phantomarbeiten bei Einsatz eines strukturierten Bewertungsbogens

Kellersmann, Christian Tim

Das Ziel vorklinischer zahnmedizinischer Behandlungskurse am Phantompatienten ist es, den Studierenden die für die Behandlung „echter“ Patienten im klinischen Studienabschnitt erforderliche kognitive und psychomotorische Kompetenz zu vermitteln. In diesem Zusammenhang müssen die Studierenden z.B. verschiedene Arten von Zahnersatz für ihren Phantompatienten anfertigen und eingliedern, wobei letztendlich die Ergebnisqualität benotet wird. Hierbei stellt sich naturgemäß die Frage nach der Verlässlichkeit der Benotung im Hinblick auf die Objektivität und Reproduzierbarkeit (inter- bzw. intraindividuelle Reliabilität).

Da bisherige Studien gezeigt haben, dass inter- und intraindividuelle Reliabilität bei der Bewertung der Ergebnisqualität restaurativer Arbeiten nur unbefriedigend ist, wenn nach dem bisher allgemein üblichen „glance and grade“-System vorgegangen wird, ist es das Ziel der vorliegenden Studie festzustellen, inwieweit der Einsatz eines strukturierten Bewertungsbogens mit definierten Bewertungskriterien die Reliabilität der Benotung steigert.

Zu diesem Zweck wurden 30 im Phantomkurs der Zahnersatzkunde angefertigte Brücken von den Studierenden selber, einem Kommilitonen, zwei zahnmedizinischen Studenten des klinischen Studienabschnittes sowie jeweils zwei Zahnärzten aus der vorklinischen und der klinischen Kursbetreuung unabhängig voneinander und wiederholt beurteilt. Die Benotung der Arbeiten erfolgte dabei sowohl mittels Checkliste als auch unter Zuhilfenahme eines strukturierten Bewertungsbogens mit vorgegebenen Beurteilungskriterien.

Durch den Einsatz des detaillierten Bewertungsbogens konnte sowohl die Objektivität (interindividuelle Reliabilität) als auch die Reproduzierbarkeit (intraindividuelle Reliabilität) der Benotung nachweisbar gesteigert werden. Dabei wies die Gruppe der klinischen Kursassistenten sowohl beim inter- als auch beim intrapersonellen Vergleich die höchste Urteilskonkordanz auf. Die geringste interpersonelle Urteilkonkordanz wies die Gruppe der Studierenden selbst auf. Wich die Selbsteinschätzung der Studierenden ohne Bewertungsbogen bis zu eine Note von der Fremdbeurteilung ab, so konnte die Konkordanz mittels des Bewertungsbogens statistisch signifikant gesteigert werden.

Die vorliegenden Ergebnisse zeigen somit, dass durch den Einsatz eines strukturierten Bewertungsbogens die Zuverlässigkeit (Objektivität und Reproduzierbarkeit) der Benotung zahnmedizinischer Phantomarbeiten signifikant gesteigert werden kann und damit über den in vergleichbaren internationalen Studien angegebenen Korrelationswerten für andere Bewertungsverfahren liegt.

Tag der mündlichen Prüfung: 07.Februar 2008

Meiner Freundin.

Inhaltsverzeichnis

1 Einleitung 1

1.1 Einführung 1

1.2 Ziel der Untersuchung 3

2 Literaturübersicht 4

2.1 Überblick zu bisherigen Studien zur Reliabilität der Bewertung vorklinischer, studentischer Phantomarbeiten 4

2.2 Einflußfaktoren bezüglich der Reliabilität der Bewertung vorklinischer Studentenarbeiten 13

2.2.1 Bewertungskriterien 13

2.2.2 Benotungssystem 15

2.2.3 Bewerter 19

3 Material und Methode 21

3.1 Bewertete Phantomkursarbeiten 21

3.2 Bewerter 21

3.3 Vorgehen bei der Bewertung 22

3.4 Statistische Auswertung 27

4 Ergebnisse 29

4.1 Vergleich der Durchschnittsnoten verschiedener Bewertergruppen bei der Benotung mittels Bewertungsbogen 29

4.2 Vergleich der Notendifferenzen innerhalb der jeweiligen Bewertergruppen bei der Benotung mittels Bewertungsbogen 31

4.3 Vergleich der studentischen Selbsteinschätzung mit dem Urteil anderer Bewerter bei der Benotung mittels Bewertungsbogen 33

4.4 Interpersonelle Urteilskonkordanz innerhalb verschiedener Bewertergruppen bei der Benotung mittels Bewertungsbogen 34

4.5 Intrapersonelle Urteilskonkordanz bei der Benotung mittels Bewertungsbogen 36

4.6 Vergleich der Bewertungen mit und ohne Bewertungsbogen 36

5 Diskussion 41

6 Zusammenfassung 45

7 Literaturverzeichnis 47

Lebenslauf 53

1 Einleitung

1.1 Einführung

In der Zahnheilkunde allgemein und speziell in der Lehre spielt die Beurteilung der Qualität von Restaurationen eine wichtige Rolle. Zur Leistungsmessung und –beurteilung in den universitären Kursen werden den Studenten genaue Vorgaben an Qualitätskriterien für die zu erbringenden Arbeiten gegeben. In den meisten Fällen kommt es bei der Bewertung dieser Arbeiten durch den subjektiven Ermessensspielraum der einzelnen Bewerter allerdings zu einer gewissen Notenstreuung [5, 6, 7, 8, 21, 34]. Dabei spielen die zur Bewertung herangezogenen Qualitätskriterien eine entscheidende Rolle. Welche dazu zählen und in welchem Maße diese in die individuelle Wertung mit einfließen, hängt von mehreren Faktoren ab. Eine unterschiedliche Ausbildung, unterschiedliche Erfahrung, ein unterschiedlicher Fortbildungsstand oder auch verschiedene zur Bewertung herangezogene Qualitätskriterien bilden einen individuellen Qualitätsanspruch.

Das Ziel vorklinischer zahnmedizinischer Behandlungskurse am Phantompatienten ist es, den Studierenden die für die Behandlung „echter“ Patienten im klinischen Studienabschnitt erforderliche kognitive und psychomotorische Kompetenz zu vermitteln. In diesem Zusammenhang müssen die Studierenden z.B. verschiedene Arten von Zahnersatz für ihren Phantompatienten anfertigen und eingliedern, wobei letztendlich die Ergebnisqualität benotet wird. Hierbei stellt sich naturgemäß die Frage nach der Verlässlichkeit = Reliabilität der Benotung.

Klassische Gütekriterien für die Reliabilität einer Bewertung sind die *Objektivität* (interindividuelle Reliabilität) und die *Reproduzierbarkeit* (intraindividuelle Reliabilität).

Eine Bewertung gilt als *objektiv*, wenn das Ergebnis unabhängig vom Bewerter gewonnen wird. Eine Notwendige Vorraussetzung für diese Unabhängigkeit sind standardisierte Prüfbedingungen. Zu unterscheiden sind Durchführungs-, Auswertungs- und Interpretationsobjektivität. Eine Methode zur Überprüfung der Objektivität lautet: Eine Bewertung ist objektiv, wenn mehrere Bewerter mit dem gleichen Bewertungssystem bei der gleichen Untersuchungspopulation möglichst gleiche Bewertungen erzielen.

Bei der *Reproduzierbarkeit* (intraindividuelle Reliabilität) geht es um die Wiederholung und damit Zuverlässigkeit der Datenerhebung. Eine möglichst hohe Reliabilität ist die Grundlage für eine valide Bewertung.

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, keine relevanten Teilaspekte zu übersehen oder zu vergessen. Diese Checklisten enthalten allerdings keine Bewertungskriterien, sondern lediglich 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 der Sicht des Bewertenden allgemein gültigen Qualitätskriterien).

Da bisherige Studien zur Reliabilität vorklinischer restaurativer Arbeiten allerdings nur eine unbefriedigende Objektivität und Reproduzierbarkeit des bisher allgemein üblichen „glance and grade“-Systems mittels Checkliste gezeigt haben, ergibt sich die dringende Notwendigkeit einer Verbesserung des bisherigen Bewertungssystems.

1.2 Ziel der Untersuchung

Vor diesem Hintergrund ist es das Ziel der vorliegenden Studie, festzustellen, inwieweit der Einsatz eines strukturierten Bewertungsbogens mit definierten Bewertungskriterien die Reliabilität der Benotung im Vergleich zum herkömmlichen „glance and grade"-System (d.h. Gesamtnote nach Zuaugenscheinnahme ohne Bewertungsbogen) mittels Checkliste bei der Benotung zahnärztlich-prothetischer Arbeiten im vorklinischen Phantomkurs steigern kann.

2 Literaturübersicht

2.1 Überblick zu bisherigen Studien zur Reliabilität der Bewertung vorklinischer, studentischer Phantomarbeiten

Die Reliabilität der Bewertung vorklinischer, studentischer Phantomarbeiten betreffend, gibt es bisher nur wenige Untersuchungen. Insgesamt waren bei einer systematischen Literaturrecherche nur 25 Untersuchungen zum Thema Reliabilität bei der Bewertung von studentischen Arbeiten zu finden (siehe Tab.1). Davon wiederum ist nur die Veröffentlichung von *Türp et al*. [34] aus dem Jahre 2002 aus dem deutschsprachigen Raum. Alle anderen Untersuchungen stammen aus dem angloamerikanischen Raum und beziehen sich damit auf andere Ausbildungssysteme.

Türp et al. [34] untersuchten die „Variabilität bei der Benotung studentischer Arbeiten im vorklinischen Phantomkurs". Im Rahmen dieser Untersuchung wurden 20 Brücken und 20 Interimsprothesen von drei Zahnärzten und drei klinischen Studenten auf das Ziel hin analysiert, das Ausmaß der interindividuellen Variabilität bei der Bewertung dieser zwei zahnärztlich-zahntechnischen Leistungen herauszustellen. Im Detail wurden die Notenbandbreite, die Strenge, die Abhängigkeit der Notenstreuung zwischen den Untersuchern und der Einfluss einer Note auf die andere untersucht.

Für die Brückenarbeit wurden sechs Teilnoten (Gesamteindruck, Sägemodell, Artikulator, Präparation, Modellation, okklusale Gestaltung, Verblendung, Ausarbeitung/Politur, Okklusion) und für die Interimsprothese fünf Teilnoten (Präparation, Klammern, Aufstellung Okklusion, Ausarbeitung/Politur) vergeben. Die Notenskala reichte von sehr gut (1) bis schlecht (6). Die anonymisierten Arbeiten wurden zufällig ausgesucht und die Probanden bewerteten unabhängig voneinander.

Die Untersuchung ergab, dass die Zahnärzte bei der Brücke im Durchschnitt um 0,47 Notenpunkte und bei der Interimsprothese um 0,42 Notenpunkte strenger bewerteten als die klinischen Studenten.

Autor	Bewertete Studentenarbeit	Reliabilität	
		Intraindiv. Reliabilität	Interindiv. Reliabilität
Bedi et al. (1987)	Klasse I-, II-, III- und Kronen-Präparationen	-	κ = 0,32 - 0,72
Dhuru et al. (1978)	Klasse II-Präparationen	-	ICC= 0,23-0,67
Feil et al. (1982)	Klasse II-AgAm, VMK-Kronen	ICC= 0,53-0,68	ICC= 0,68-0,8
Fuller et al. (1972)	Klasse II-Präparationen	r_s= 0,47-0,83	r_s= 0,2-0,56
Gaines et al. (1974)	Kronen-modellationen	-	ICC= 0,26-0,56
Goepferd et al. (1980)	Klasse II-Präparationen	r_p= 0,62-0,68	ICC= 0,3-0,47
Hinkelman et al. (1973)	Kavitätenpräparation, Füllungen, Inlays	-	46,7%-84%
Houpt et al. (1973)	Klasse II-Präparationen	r_p= 0,36-0,63	r_{Fin}= 0,61-0,75
Lilley et al. (1968)	Amalgam-füllungen	r_p= 0,51-0,63	r_p= 0,11-0,72
Meetz et al. (1988)	technical skills	-	r_p= 0,62-0,83
Natkin et al. (1967)	Wurzelkanal-behandlungen	-	Ø Notenabweichung 1,2-1,48
Robertello et al. (1997)	Amalgam-füllungen	83-92%	61-70%
Türp et al. (2002)	Verblendbrücken partielle Prothese	-	ICC= 0,61
Vann et al. (1983)	Klasse-II-Kavitäten	r_p= 0,46-0,86	ICC= 0,32-0,49

Tab.1 *Ergebnisse bisheriger Untersuchungen zur inter- und intraindividuellen Reliabilität bei der Bewertung von Phantomkursarbeiten.*

Der in den klinischen Kursen eingesetzte Zahnarzt bewertete am strengsten. Des Weiteren konnte festgestellt werden, dass bei einer Note von ≤ 3,5 für die Brücken-Bewertung eine überproportionale Wahrscheinlichkeit bestand, auch für die Interims-prothese eine Note von ≤ 3,5 zu erhalten.

Was den Gegenstand bisheriger Untersuchungen zur Reliabilitätssteigerung bei der Bewertung von Phantomkursarbeiten betrifft, so wurden fast ausschließlich konservierende Maßnahmen wie Klasse II Präparationen und/oder Füllungen beziehungsweise

endodontische Maßnahmen bewertet (Tab. 1). Lediglich die Untersuchung von *Türp et al*. [34] bezog sich auf prothetische Arbeiten des vorklinischen Phantomkurses.

Welchen Einfluss der Gebrauch von Kriterien beziehungsweise Checklisten auf die Reliabilität der Bewertung hat, untersuchten unter anderem *Vann et al*. [36] und bezogen sich dabei unter anderem auf *Goepferds und Kerbers* [9] Studie, welche einen Vergleich zwischen einer globalen Bewertungsmethode (Bewertung des Ganzen) und einer analytischen Methode (Bewertung von Teilleistungen nach bestimmten Kriterien einer Checkliste) beschrieb. Hierbei zeigte sich, dass inter- und intraindividuelle Reliabilität bei der analytischen Methode etwas höher waren, wobei der Unterschied jedoch nicht statistisch signifikant war.

In der Studie von *Vann et al.* [36] wurden folgende drei Bewertungsmethoden untersucht:

1. **„Glance and grade“**-System:
   Bewertung der Gesamtleistung, keine genauen Kriterien oder Checklisten; Noten A,B,C,D,F
2. **Checklisten-Methode**:
   Liste mit zehn Unterpunkten für die Kavitätenbewertung, jeder Unterpunkt erhält einen Wert von 1 (schlecht) bis 5 (sehr gut)
3. **Analytische Methode**:
   jeder Unterpunkt ist zusätzlich mit einer Kriterienliste versehen, pro Kriterium 1-5 Punkte; Methode identisch mit *Goepferds und Kerbers* [9] analytischer Methode

Der Versuchsaufbau bei *Van et al.* [36] bestand aus drei Studenten und drei Assistenzzahnärzten, die jeweils 30 Klasse II-Präparationen bewerteten. Die Auswahl der Präparationen erfolgte randomisiert aus 83 Modellen einer Arbeitsprobe. Um Übungseffekte zu vermeiden, bewerteten alle Bewerter in unterschiedlicher Reihenfolge. Jeweils nach 24 Stunden erfolgte eine erneute Bewertung der Modelle mit gleicher Methode, jedoch in geänderter Reihenfolge, was zur Untersuchung der intrinsischen Reliabilität diente. Die Bewerter wurden bei jeder Methode durch Handzettel und Anweisungen auf die Bewertungen vorbereitet. Zusätzlich wurde ihnen ein Überblick über die Untersuch-

ungsziele vermittelt. Bei den Untersuchungen von *Goepferd und Kerber* [9] wurden die Probanden hingegen nur bei der analytischen Methode vorbereitetet.

Auch *Feil et al*. [6] haben sich mit der Problemstellung der Bewertung studentischer Arbeiten beschäftigt. Sie sahen als Hauptursachen für Bewertungsfehler den Bewerter an sich, unklare Bewertungskriterien und die variable Bewertungsbreite. Sie versuchten die Reliabilität zu steigern, indem sie die Bewertungsmängeln am herkömmlichen „glance and grade"-System wie die fehlende Spezifizierung durch subjektive Einschätzungen, ein nicht ausreichendes Studenten-Feedback sowie ein zu großer Zeitaufwand für die Notenberechnung optimierten.
Das von ihnen entwickelte neue Bewertungs-System basierte auf einem Bewertungsbogen, der verschiedene Bewertungen einzelner Kriterien vorgab. Effizientes Studenten-Feedback sollte dadurch erzielt werden, dass die Studenten zuerst ihre eigene Arbeit mit dem Bogen evaluierten, bevor die Ausbilder dies taten. Hierdurch sollte die Notengebung für den Studenten transparenter gestaltet werden.
Feils Reliabilitätsstudie nahm ihre Daten aus zwei regulären zahntechnischen Kursen in denen sowohl eine VMK-Krone und eine mod-Kavität präpariert als auch eine mod-Restauration gelegt wurden. Zufällig gepaarte Bewerter wurden zufällig zu Studenten-Gruppen zugeordnet und jeder Bewerter beurteilte jeden Studenten.
Der durch den Bewertungsbogen erzielte Intraklassenkorrelationskoefizient (ICC) lag bei der mod-Kavität bei 0.80, bei der mod-Restauration bei 0.74 und bei der VMK-Kronenpräparation bei 0.68.

Bereits 1973 hatten sich *Hinkelmann und Long* [13] mit ähnlichen Problemen in der Bewertung von zahnärztlichen Arbeiten beschäftigt.
Grundlage ihrer Untersuchung war die Erarbeitung eines Bewertungsbogens, mit dem sowohl die Präparationsphase als auch die Restaurationsphase einer vorklinisch geleisteten Arbeit bewertet werden sollte. Das Evaluationssystem, welches Einzelkomponenten jeder Phase und die Gesamtleistung einbezieht, basierte dabei auf einem Drei-Punkt-System. Nach Autorenmeinung würde ein Zwei-Punkt-System die Studenten dazu verleiten, nur den minimalen Aufwand zu leisten, der notwenig war, um die Arbeit zu bestehen. Folglich erhielt eine Arbeit, die keine Verbesserungen benötigt, drei

Punkte, eine klinisch akzeptable Leistung zwei und eine klinisch unakzeptable, unkorrekte Leistung null Punkte. Der Durchschnittspunktwert aller Einzelkategorien bestimmte anschließend das Endresultat. Eine einzige Bewertung von null Punkten reduzierte das Gesamtergebnis bereits unter das Bestehensniveau.
Die im Testbogen berücksichtigten Kategorien stammten teils aus einem Lehrbuch, teils aus Anregungen der Kursassistenten. Nachdem der Bogen das gesamte Semester durch für jede Bewertung benutzt wurde, überprüften die Autoren zum Kursende dessen Reliabilität. Hierzu dienten 60 zufällig ausgewählte und anonymisierte Inlay-Präparationen, welche von vier Assistenten bewertet wurden. Zwei Assistenten hatten sieben Jahre, die anderen beiden weniger als ein Jahr Berufserfahrung. Sie bewerteten die ersten 30 Prä-parationen unabhängig voneinander, vier Wochen später die verbleibenden 30. Zur Auswertung wurde kein Reliabilitätskoeffizient benutzt, da dieser bei einem Drei-Punkt-System irreführende Ergebnisse produzieren würde. Daher wurde lediglich die prozentuale Übereinstimmung der Bewerter bestimmt.
Bei 56,3% lag das Ergebnis vor und bei 58% Übereinstimmung das Ergebnis nach einer Trainingsdiskussion.

Fast zur selben Zeit, im Jahr 1972 untersuchte *Fuller* [7] die Übereinstimmung von Bewertern mit herkömmlicher Bewertungsmethode sowie die Konstanz der Bewertung herauszustellen. Darüber hinaus sollte nachgewiesen werden, inwiefern die Übereinstimmung der Bewertung durch Training oder andere Methoden gesteigert werden kann, und ob die Sequenz des Trainings dabei einen Einfluss auf das Ergebnis hat.
Vier Bewerterpaare bewerteten die Leistungen von drei praktischen Prüfungen der 67 Erstsemester-Studenten der Universität von Iowa. Art der Leistung waren Präparationen vorgegebener Kavitäten. In der ersten Bewertungsrunde bewerteten alle Versuchspersonen separat die anonymisierten Arbeiten durch kurze Inspektion, kurz darauf noch einmal 25 zufällig ausgewählte. Dies diente dazu, die interindividuelle Reliabilität der Bewerter festzustellen. Danach wurden Paarungen gebildet. Paar 1 bewertete auch Durchgang zwei und drei nach der herkömmlichen Methode, Paar 2 bewertete Durchgang zwei herkömmlich und Durchgang drei mit Checkliste + Training, Paar 3 bewertete Durchgang 2 mit Checkliste und Durchgang drei nach einem Training, Paar vier bewertete Durchgang 2 nach absolviertem Training und Durchgang drei mit der

Checkliste. Die Checkliste umfasste Überpunkte wie okklusale Form und approximale Form für die Klasse II Amalgam-Füllung, gesamte Extensionsform, innere Kavitäten-form und Retentionsform für die Klasse III Gold-Präparation. Diese Überpunkte waren in bis zu acht Unterpunkte aufgeschlüsselt. Im Trainingsprogramm wurden 1. die Lehrinhalte/Lehrtechniken erklärt und diskutiert, 2. Faktoren wie Sympathie, Rahmen-bedingungen, bewerterspezifische Gewichtung einzelner Kriterien besprochen, 3. die Benotungstechnik erklärt und 4. eine praktische Bewertungsübung unter Aufsicht durchgeführt. Die Benotungstechnik basierte auf einem Punktsystem von 1-50. Alle Arbeiten wurden zunächst in fünf Gruppen eingeteilt. Die beste Gruppe erhielt beispielsweise Punktwerte von 41-50, die schlechteste von 1-10.
Hierbei ergab sich die höchste intraindividuelle Übereinstimmung mit einem ICC von 0.831 beim erfahrensten Assistenten und die geringste intraindividuelle Übereinstimmung mit einem ICC von 0,663 beziehungsweise 0,472 bei den unerfahrensten Assistenten mit nur einem Jahr Lehrerfahrung. Die intraindividuellen ICCs der verschiedenen Paarungen zeigten deutlich, dass weder die Checkliste, das Training, noch beides in Kombination zur gewünschten Reliabilitätssteigerung geführt hatten.

Bereits Jahre zuvor untersuchten *Lilley et al.* [21], ob durch Festlegung notenwirksamer Kriterien und genauer Fehlerdefinition die intraindividuelle sowie interindividuelle Reliabilität gesteigert werden kann.
Hierzu führten sie eine Untersuchung durch, bei der 37 anonymisierte Modelle mit Studentenarbeiten von drei Probanden bewertet wurden. Die Modelle wiesen jeweils vier Kunststoffzähne auf, an denen die vier Stufen einer Amalgamfüllung (Präparation, Füllung, Ausarbeitung, Politur) ausgeführt wurden. So konnten die Bewerter danach alle Stufen gleichzeitig und wiederholbar bewerten. Bewertungen von A bis E wurden jeder einzelnen Stufe und auch der Gesamtarbeit zugeteilt. Die Bestehensgrenze lag zwischen C und D. Einen Monat nach dem Bewertungsdurchgang wurden die Modelle erneut bewertet. Zuvor diskutierten die Probanden Bewertungskriterien für alle Stufen und es wurden Fehler aufgezeigt, die zum sofortigen durchfallen führten. Nach einem weiteren Monat fand ein dritter Bewertungsdurchgang statt.
Die interindividuelle Variabilität war, sofern nur die Bewertung der Kavität betrachtet wurden, im ersten Durchgang sehr hoch (ICC 0,12 - 0,28). Durch die Diskussion der

Bewertungskriterien vor dem zweiten und dritten Bewertungsdurchgang verringerte sie sich jedoch deutlich, so dass Korrelationskoeffizienten (ICC) von 0,37-0,43 (zweiter Durchgang) beziehungsweise 0,39-053 (dritter Durchgang) gemessen wurden. Die intraindividuelle Variabilität lag hingegen deutlich niedriger, hier betrug der Korrelationskoeffizienten (ICC) 0,51-0,63.

Gaines et al. [8] untersuchten ebenfalls die interindividuelle Reliabilität verschiedener Bewerter. Es sollte die Konstanz zwischen Bewertern, die zwei verschieden objektive Systeme nutzen, untersucht werden. Fokussiert wurden dabei insbesondere die Unterschiede zwischen den Bewertern in den verschiedenen Bewertungsrunden sowie deren Einschätzung der Reliabilität der individuellen Bewertung.
Versuchsteilnehmer waren sieben Angehörige der prothetischen Fakultät, welche zwei Modellations-Arbeitsproben von acht zufällig ausgewählten Studenten unabhängig bewerteten. Der Untersuchung lagen die folgenden zwei Bewertungsarten zugrunde:

1. Die bisher benutzte Bewertungsmethode, bei welcher Punkte (5=hervorragend, 1=schlecht) für sechs Teilbereiche der Modellation vergeben wurden.
2. Die modifizierte Bewertungsmethode, bei welcher den Punktwerten eine bildliche Beschreibung der Leistung zugeordnet wurde. So wurde beispielsweise die Ausarbeitung mit Adjektiven wie „vollkommen glatt“, „leicht rau“ oder „grobe Kratzer“ beschrieben.

Es wurde ein Bewertungsdurchgang mit jeder Methode durchgeführt. Um Übungseffekte zu verhindern, wurden in der zweiten Runde acht andere Arbeiten als in der ersten Runde bewertet.
Nach Datenauswertung ergaben sich statistisch signifikante Unterschiede zwischen den Bewertern bei der Bewertung gleicher Arbeiten. Der maximale Unterschied betrug 22,6 Punkte, bei einem Durchschnittspunktwert von 31,63 zu 54,25. Der Korrelationskoeffizient der ersten Bewertung lag bei 0,26, der der zweiten Bewertung bei 0,56. Die Autoren schlussfolgerten, dass deskriptive Punktverteilung eine größere Konstanz mit weniger ausgeprägten Varianzen liefert.

Natkin und Guilds [26] bewerteten 65 endodontische Studenten-Behandlungen extrahierter Zähne, hierbei vergaben sechs Bewerter neun mögliche (Teil-)Noten mit erklärendem Kommentar.
Nach klassischer Bewertung („glance and grade“) zeigten sich deutliche Unterschiede in der Notengebung: 45% der Zähne erhielten für dieselbe Arbeit Noten mit einer Bandbreite von vier Noten Unterschied. So wurde z.B. eine Arbeit von Bewerter 3 mit „gut“ und von Bewerter 4 mit „unmöglich“ bewertet. Nur bei 8% der untersuchten Arbeiten betrug die Notendifferenz weniger als eine Note.

Eine ähnliche Studie mit Probanden unterschiedlicher Lehrerfahrung führten auch *Jenkins et al.* [17] im Jahr 1998 durch. Bewertet wurden 75 Klasse II Präparationen für Amalgamfüllungen, von denen 54 von vorklinischen Studenten und 21 von einem Vollzeit-Assistenzarzt stammten. Es wurden zwei separate Bewertungszyklen unter standardisierten Bedingungen durchgeführt. Als Untersuchungshilfe standen zahnärztliche Sonden und Parodontalsonden mit Millimeter-Skalierung zur Verfügung. Die Probanden erhielten schriftliche Informationen über die Idealpräparation, jedoch keine detaillierte Checkliste mit Kriterien oder potentiellen Fehlern. Es konnten insgesamt 13 (Teil-) Noten vergeben werden, welche zur Auswertung in Punkte umgewandelt wurden.
In Bezug auf die intraindividuelle Reliabilität wiesen sowohl der erfahrenste als auch der unerfahrenste Bewerter Abweichungen von weniger als fünf Punkten auf. Bei den übrigen Bewertern lagen größere Differenzen von bis zu 7 und mehr Punkten vor.
Zur Darstellung der Urteilskonkordanz zwischen den Bewertern wurden alle Ergebnisse mit denen von Bewerter 1 verglichen. Auch hier ergaben sich Unterschiede von mehr als fünf Punkten. Die Gruppe der Vollzeit-Assistenten, d.h. derjenigen mit der längsten Berufserfahrung, vergaben insgesamt die besten Noten.

In Anlehnung an eine Studie von *King und Bedi* [19] aus dem Jahre 1984, welche bildliche Bewertungskriterien für die Evaluation von Milchzahnrestaurationen vorschlug, untersuchten *Bedi et al.* 1987 [3] den Einfluss bildlicher Kriterien auf die Reliabilität von Bewertungen. Die Probanden setzten sich aus einer Gruppe mit fünf Zahnärzten, welche in der Lehre tätig waren, aus einer Gruppe mit 14 Examenskandidaten und aus

einer Gruppe mit 16 Erstsemesterteilnehmern zusammen. Sie alle bewerteten fünf verschiedene Modelle mit jeweils fünf Kavitätenpräparationen in Milchzähnen. Es fanden drei Bewertungsdurchgänge in wöchentlichem Abstand mit dem herkömmlichen Schulnoten-System statt. Danach wurden den Schulnoten bildliche Bewertungskriterien zugeordnet, die den Probanden detailliert erklärt wurden. Es folgten drei weitere Bewertungsdurchgänge.

Nach statistischer Auswertung der Untersuchungsergebnisse zeigte sich für alle Probandengruppen bei Nutzung der bildlichen Kriterien eine sehr hohe interindividuelle Reliabilität mit einem Kappakoeffizienten von 0,71. Die auffälligste Reliabilitätssteigerung durch Einsatz der bildlichen Kriterien war in der Gruppe der Examenskandidaten festzustellen. Die Notengebung der Examenskandidaten war mit einer Durchschnittsnote von 3,4 besser als die der Zahnärzte mit 3,71. Insgesamt lag die Durchschnittsnote bei Nutzung der bildlichen Kriterien mit 3,76 höher als bei Nutzung der herkömmlichen Methode (3,55). Basierend auf diesen Ergebnissen äußerten die Autoren die Vermutung, dass die bildlichen Bewertungskriterien eine positive Auswirkung auf das studentische Lernverhalten haben, da hierdurch angeblich die Selbsteinschätzung der Studenten stark erleichtert würde.

Dhuru et al. [5] berichteten 1978 über ein an Kriterien orientiertes Bewertungssystem. 52 Präparationen für verschiedenartige Füllungen in Kunststoffzähnen wurden von zwei unterschiedlich erfahrenen Bewertergruppen untersucht. Die erfahrene Gruppe bestand aus acht Zahnärzten, welche 1-9 Jahre Lehrerfahrung aufwiesen. Die vier Zahnärzte der unerfahrenen Gruppe verfügten lediglich über zwei Monate Lehrerfahrung. In der ersten Phase bewerteten sie alle Präparationen auf einem 10-Punkte-Bogen nach ihren eigenen klinischen Maßstäben. In der zweiten Phase wurden zehn zu bewertende Punkte festgelegt und genau definiert, wie sie erfüllt sein müssen, um eine positive Bewertung dafür zu erlangen. Ein nicht erfüllter Punkt wurde nicht gewertet, folglich erhielt eine perfekte Arbeit zehn Punkte, waren zwei Kriterien nicht erfüllt erhielt sie acht Punkte usw. Zur Messung der Reliabilität wurde der ICC herangezogen und Unterschiede zwischen beiden Bewertungsphasen auf ihre statistische Signifikanz untersucht.

Hierbei ergab sich, dass der Korrelationskoeffizient durch das kriterienorientierte Bewertungssystem mit vordefinierten Bewertungspunkten in der erfahrenen Gruppe um

insgesamt 0,131 gesteigert wurde. In der unerfahrenen Gruppe lag die Steigerung bei durchschnittlich 0,072. Der größte Anstieg war in beiden Gruppen bei der Bewertung der Inlay-Präparation zu vermerken.

2.2 Einflußfaktoren bezüglich der Reliabilität der Bewertung vorklinischer Studentenarbeiten

Alle vorliegenden Veröffentlichungen, unabhängig vom Gegenstand der Untersuchung, kommen in ihrem Fazit dazu, dass grundsätzlich nur eine geringe Reliabilität bei der Bewertung von studentischen Arbeiten vorliegt. Sie blieb in verschiedenen Studien inkonstant im Bereich von r_s=0,2-0,83 bzw. ICC=0,23-0,68.
Dennoch konnte die Reliabilität zum Beispiel durch die genaue Definition von Bewertungskriterien sowie deren Diskussion im Vorfeld der Untersuchung gesteigert werden. Eine signifikante Reliabilitätssteigerung konnte zwar noch durch keine Methode erzielt werden, zusammenfassend lässt sich aber sagen, dass die Reliabilität unter anderem von drei Einflussfaktoren abhängig ist:

- Beurteilungskriterien
- Benotungssystem
- Bewerter

Dies soll im Folgenden detailliert beschrieben werden.

2.2.1 Bewertungskriterien

Bewertungskriterien geben dem Bewerter vor, nach welchen Maßgaben die studentische Arbeit zu bewerten ist. Je nachdem inwieweit die vorgegebenen Bewertungskriterien den subjektiven Ermessensspielraums des Bewerters einschränken, kann zwischen drei Abstufungen unterschieden werden.

Den weit aus größten Spielraum bieten die so genannten einfachen Kriterien wie Paßgenauigkeit, Randschluß etc. Diese einfachen Kriterien ohne eine nähere Erläuterung lassen den Bewertern subjektiven Ermessensspielraum, so dass es je nach Ausbildung und eigener Neigung eine große Notenbandbreite unter und innerhalb der Bewerter geben kann. Eine weitere Einschränkung in der subjektiven Bewertung kann durch eine einfache Skalierung der Kriterien von z.B. „gut" bis „schlecht" erzielt werden. Aber erst die genaue Beschreibung der einzelnen Skalierungen die das Qualitätsniveau der Kriterien exakt definieren, z.B. der Randschluss ist mit „gut" zu bewerten, wenn der Randspalt im Bereich von 50-100 µm liegt etc., führt zu größtmöglichen Einschränkung des Ermessenspielraums.
Diese Art von vordefinierten Kriterien soll somit die Diskrepanz, die abhängig von der Subjektivität der Bewerter ist, minimieren.

In allen bereits durchgeführten Untersuchungen wird den Bewertungskriterien eine hohe Bedeutung bei der Reliabilitätssteigerung beigemessen (siehe Tab 2). In der Literatur wird der Erfolg allerdings kontrovers beschrieben.
So kommt *Bedi et al.* [3] in ihrer Veröffentlichung zur Erkenntnis, dass Bewerter deutlich höhere Kappa-Werte bei der Verwendung von definierten Kriterien im Vergleich zur globalen Methode zeigten. Bei Klasse I-, II-, III- und Kronenpräparationen stieg der Kappa-Wert bei der interindividuellen Reliabilität von K=0,52 auf K=0,72.
Gaines [8] stellte fest, dass durch einen definierten Bewertungsbogen im Vergleich zu einer Checkliste die intraindividuelle Reliabilität von einem ICC=0,26 auf 0,56 verbessert und auch die interindividuelle Urteilskonkordanz erheblich gesteigert wurde.
Des Weiteren zeigten auch die Untersuchungen von *Robertello* [29], dass ein Bewertungsbogen mit definierten Kriterien die Reliabilität der Bewerter steigern kann.
Andere Untersucher wie *Fuller* [7] äußerten sich eher pessimistisch, da sie keine signifikante Steigerung der Reliabilität durch definierte Kriterien feststellen konnten, allerdings stieg der ICC von 0,4 auf 0,58.
Zu dem gleichen Ergebnis kamen *Vann et al.* [36], die bei Bewertungen von Klasse-II-Kavitäten durch die Globale Methode, eine Checkliste und einem Bewertungsbogen keine Steigerung der inter- und intraindividuellen Reliabilität feststellten.

Autor	Arbeit	ohne Bewertungsbogen		mit Bewertungsbogen	
		intraindiv. Reliabilität	interindiv. Reliabilität	intraindiv. Reliabilität	interindiv. Reliabilität
Bedi et al. (1987)	Klasse I-, II-, III- und Kronen-Präparationen	-	$\kappa = 0,32$	-	$\kappa = 0,72$
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	ICC = 0,74
Fuller et al. (1972)	Klasse II-Präparationen	$r_s = 0,72$	ICC = 0,4	-	ICC = 0,38
Gaines et al. (1974)	Kronen-modellationen	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)	endodontische Maßnahmen	-	Ø Notenabw. 4,16	-	Ø Notenabw. 3,34
Robertello et al. (1997)	Amalgam-füllungen	83%	61%	92%	70%
Türp et al. (2002)	Verblendbrücken partielle Prothese	-	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 *Einfluss der Definition von Bewertungskriterien auf die Reliabilität der Bewertung von Phantomarbeiten.*

2.2.2 Benotungssystem

Beim Benotungssystem unterscheidet man drei Arten:

- globales System
- Checkliste
- definierter Bewertungsbogen

Bei dem so genannten **globalen System** oder auch **„glance and grade"-System** bewertet der Prüfer nach seinen eigenen Erfahrungswerten und persönlich gesetzten Schwerpunkten ohne Vorgaben z.B. in Form eines Bewertungsbogens.

Die Bewertungskriterien, die er beim „kurzen Hinsehen und Bewerten“ anbringt, liegen im Ermessen des Bewerters. Dabei spielt die individuelle Aus- und Weiterbildung die entscheidende Rolle bei der Gewichtung seiner Bewertungskriterien.
Daraus resultiert naturgemäß sowohl eine hohe Variabilität unter zwei oder mehreren Bewertern als auch bei ein und demselben Bewerter.

Die **Checkliste** soll dem Bewerter als Leitfaden bei der Bewertung von zahnärztlich, technischen Arbeiten dienen, um die Variabilität bei der Leistungsmessung und –beurteilung durch Bewertungskriterien zu minimieren, indem die zu bewertenden Aspekte einer Arbeit und eine Notenskala vorgegeben werden. Die ersten Ansätze dieser Methode gehen bereits auf *Lilley et al.* [21] zurück, der im Jahr 1968 erstmals die Intra- und Intervariabilität bei der Bewertung von technischen Arbeiten untersuchte. Die Bewerter nutzten einen Fünf-Punkte-Plan von exzellent bis hoffnungslos. Dabei wurden Arbeiten von 37 Studenten und zwar jeweils eine disto-okklusale Präparation, eine Unterfüllung, eine Amalgamfüllung sowie eine Politur von drei Untersuchern unabhängig voneinander bewertet. In monatlichen Intervallen wurden dieselben Arbeiten ein zweites und drittes Mal nach vorheriger Diskussion, in der Aspekte für die fünf Noten festgelegt wurden, bewertet. Die Ergebnisse zeigten, dass die Fünf-Punkte-Skala zu einer höheren Intervariabilität geführt hat, als die Bewertung ohne Skala. Durch das vorherige Training wurde die Intervariabilität nur gering gesenkt. Die Intravariabilität fiel geringer aus als die Intervariabilität.
Diese Untersuchung zeigte, dass eine allgemeine Bewertungsaufteilung von „exzellent“ bis „hoffnungslos“ immer noch den gleichen subjektiven, individuellen Ermessensspielraum für den Prüfer lässt, wie das „glance and grade“-System.

Folglich versuchte man nicht eine Aufteilung des Notenspektrums vorzugeben, sondern konzentrierte sich mehr auf die Definition von Teilaspekten der jeweiligen Arbeit, die einzeln bewertet wurden.
Fuller [7] gab den Prüfern bei seinen Untersuchungen zur Beurteilung von Klasse II Kavitäten neben der „glance and grade“-System zusätzlich eine Checkliste vor. Diese beinhaltete mehrere Kriterien für die Bewertung der okklusalen und approximalen Form. Bei der Versuchsdurchführung bewerteten vier Bewerterpaare alle praktischen

Arbeiten der Erstsemesterstudenten der University of Iowa. Dabei bewertete ein Paar während der gesamten Studie mittels der traditionellen „glance and grade“-System, während die anderen drei Paare verschiedene Kombinationen von Training und der Checkliste anwendeten. Ziel der Studie war es sowohl die Inter- sowie Intrareliabilität als auch den Faktor Training der Checkliste zu untersuchen. Hierbei ergab sich, dass selbst die Bewertung durch die Checkliste zu keiner signifikanten Steigerung der Reliabilität führt. Er empfahl daher für zukünftige Untersuchungen, die Bewertungskriterien spezifischer zu gewichten und zu definieren.

Im Weiteren versuchte man daher die **Bewertungskriterien** so exakt zu definieren, dass kein subjektiver Ermessensspielraum für den Prüfer verbleibt, um so die Variabilität auf ein Minimum reduzieren zu können (siehe Tab.3).
Houpt und Kress [14] untersuchten 1973 in diesem Zusammenhang, ob die Anzahl an Bewertungspunkten und eine entsprechende Definition einen Einfluss auf die Reliabilität bei der Bewertung haben. Dazu benutzten sie drei verschiedene Bewertungsbögen, die sich in der Anzahl und der Definition von Bewertungspunkten unterschieden.
Bewertungsbogen A umfasste lediglich zwei Bewertungspunkte pro Kriterium, und zwar korrekt oder inkorrekt. Bewertungsbogen B umfasste fünf Bewertungspunkte von 0 bis 4 pro Kriterium. Beim Bewertungsbogen C wurden die Bewertungspunkte 0 bis 4 noch einmal klar definiert.
Der 2-Punkte-Bewertungsbogen ergab eine höhere Reliabilität als der 5-Punkte-Bewertungsbogen. Auch die genau definierten Bewertungspunkte unterlagen dem 2-Punkte-Bewertungsbogen. Dennoch sollte kritisch hinterfragt werden, ob ein „richtig“ und „falsch“ als alleinige Bewertungskriterien ausreichend sind, da sie die Bandbreite der Arbeiten stark einschränkt. Es stellt sich also bei jeder Untersuchung die Frage, was noch als „gerade richtig“ und was als „schon falsch“ zu bewerten ist? Das 2-Punkte-System kann dieser Anforderung nicht gerecht werden.

Durch exakt definierte Bewertungskriterien soll eine Arbeit so genau bewertet werden können, dass es gar keinen Ermessensspielraum für die Bewerter mehr gibt und so die Reliabilität gesteigert werden kann.

Gaines et al. [8] verwendeten solche exakt definierten Bewertungskriterien in ihren Untersuchungen und konnten die intraindividuelle Reliabilität von ICC=0,26 auf ICC=0,56 steigern. *Hinkelman et al*. [13] führten eine ähnliche Studie durch und konnten ebenso eine hohe Übereinstimmung unter den Bewertern feststellen.

Neben der Entwicklung von Bewertungsbögen gab es die Überlegung von *Bedi und King* [3]. eine bildhafte Form der Kriterien zur Bewertung restaurativer Arbeiten zu entwickeln

Diese **bildhaften Kriterien**, die die verschiedenen Arbeitsschritte z.B. einer Präparation wiedergeben, sollten einfach zu benutzen und zu merken sein. Das Ziel der Überlegung war, eine bessere Bewertungsform gegenüber der „glance and grade"-System für die Prüfer und eine Lernhilfe für die Studenten zu entwickeln.

Für ihre Untersuchungen stellten sie fünf verschiedene Modelle mit jeweils fünf verschiedenen Präparationen zur Verfügung. Zur Bewertung zogen sie die schon 1984 von ihnen propagierten Kriterien heran [19]. Dieses waren Zeichnungen der verschiedenen Kavitätenpräparationen, die diese optimal darstellten, aber auch Anschauungsbeispiele mit Fehlern, welche unakzeptable Präparationen zeigten.

Als Bewerter der Studie wurden fünf Assistenten, 14 Studenten im letzten Jahr und 16 Studenten im ersten klinischen Jahr nach dem Zufallsprinzip ausgewählt. Alle bewerteten die fünf Modelle drei Mal mit jeweils einer Woche Pause zwischen den Bewertungen mittels der üblichen „glance and grade"-System von A (excellent) bis E (poor).

Anschließend wurden die Durchgänge mittels der Zeichnungen als Goldstandard wiederum dreimal durchgeführt. Zur Auswertung der Reliabilität wurden die Werte in ein Computersystem eingegeben und mittels Cohen's Kappa bewertet.

Das Ergebnis zeigte eindeutig, dass die Reliabilität mit den bildhaften Kriterien wesentlich höher lag als ohne. Im Durchschnitt betrug der Kappa-Koeffizient als Reliabilitätsmaß für die Bewertung ohne bildliche Kriterien bei 0,32 und bei der Bewertung mit bildlichen Kriterien bei 0,72.

Trotz dieser unumstrittenen Verbesserung der Reliabilität im Vergleich zum „glance and grade"-System, sollte die Bewertung nach bildhaften Kriterien nach Meinung von *Bedi* und *King* selbst eher als Hilfsmittel für das Selbststudium der Studenten und nicht

als Grundlage einer soliden Bewertung betrachtet werden. Denn die Beurteilung einer Arbeit im Vergleich zu einer vorgegebenen Zeichnung biete weiteren Spielraum für subjektive Beurteilungen und führe zu keiner wissenschaftlich begründeten und nachweisbaren Bewertung.

Autor	Arbeit	unspez. Notenkriterien		def. Notenkriterien	
		intraindiv. Reliabilität	interindiv. Reliabilität	intraindiv. Reliabilität	interindiv. Reliabilität
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)	**endodontische Maßnahmen**	**-**	**Notenabw. 4,16**	**-**	**Notenabw. 3,34**
Robertello et al. (1997)	**Amalgam-füllungen**	**83%**	**61%**	**92%**	**70%**
Vann et al. (1983)	**Klasse II-Präparationen**	**r_p= 0,75**	**ICC= 0,34**	**r_p= 0,73**	**ICC= 0,33**

Tab.3 *Einfluss des Benotungssystems auf die Reliabilität bei der Bewertung von Phantomarbeiten.*

2.2.3 Bewerter

Der dritte Einflußfaktor bei der Bewertung von zahnärztlichen Arbeiten ist der Bewerter an sich. In mehreren Studien wurde versucht, die Subjektivität als Inkonstante durch Trainingsseminare zu reduzieren. Aber weder *Houpt und Kress* [14] noch *Fuller* [5] stellten eine signifikante Steigerung der Reliabilität durch Training der Bewerter fest. Bei *Houpt und Kress* [14] erhielten die Bewerter nach einem Bewertungsdurchgang ein Feedback über ihre Bewertung. Zur Kalibrierung der Bewerter geschah dies immer im direkten Vergleich zu den Ergebnissen der anderen Bewerter.

Bei *Fuller* [5] erhielten die Bewerter im Vorfeld der Evaluation ein zweistündiges Training, in dem das Bewertungssystem und seine Anwendung besprochen wurden.

Im Gegensatz dazu konnte bei der Untersuchung von *Natkin* und *Guild* [26] eine Verbesserung der Reliabilität erzielt werden. Im Rahmen ihrer Studie wurden die Bewerter durch Diskussionen in mehreren Treffen trainiert und somit gezielt auf die Bewertung vorbereitet. Neben dem Versuch, die Reliabilität durch eine Kalibrierung der

Bewerter zu steigern, wurde in der Literatur oft untersucht, ob die Bewertererfahrung Einfluss auf die Reliabilität hat.

Dazu ließ *Dhuru et al.* [5] erfahrene und unerfahrene Bewerter jeweils einmal mit und einmal ohne Bewertungsbogen studentische Arbeiten bewerten. Die geringste Reliabilität zeigten die unerfahrenen Bewerter ohne Bogen und die größte Reliabilität die erfahrenen Bewerter mit Bogen. Außerdem lag die Reliabilität der unerfahrenen Bewerter mit Bogen nur geringfügig über der Reliabilität der erfahrenen Bewerter ohne Bogen.

Dies zeigt, dass die Berufserfahrung eine entscheidende Rolle bei der Benotung von studentischen Arbeiten spielen kann (siehe Tabelle 4).

Autor	Hilfsmittel	Reliabilität, abhängig vom Grad der Bewertererfahrung					
		gering		mittel		viel	
		nein	ja	nein	ja	nein	ja
Bedi et al. (1987)	**Checkliste**	**-**	**-**	**-**	**k=0,33**	**-**	**k=0,39**
Dhuru et al. (1978)		**ICC=0,47**	**ICC=0,54**	**-**	**-**	**ICC=0,52**	**ICC=0,65**
Goepferd et al. (1980)	**Klasse II-Präparationen**	**ICC=0,30**	**ICC=0,47**	**-**	**-**	**r_p=0,62**	**r_p=0,68**
Hinkelman et al. (1973)	**Training**	**63,4%**	**49,2%**	**-**	**-**	**56,0%**	**68,0%**
Meetz et al. (1988)	**technical skills**	**r_p=0,62**	**r_p=0,83**	**-**	**-**	**-**	**-**
Natkin et al. (1967)		**Ø Notenabweichung**					
		4,16	**-**	**3,69**	**-**	**3,34**	**-**
Türp et al. (2002)		**-**	**ICC=0,61**	**-**	**-**	**-**	**-**

Tab.4 Einfluss der Bewerter auf die Reliabilität bei der Bewertung von Phantomarbeiten.

3 Material und Methode

3.1 Bewertete Phantomkursarbeiten

Aus dem Phantomkurs der Zahnersatzkunde II am Universitätsklinikum Münster wurden im Sommersemester 2003 nach Kursende 30 das gesamte Notenspektrum abdeckende repräsentative Arbeiten ausgewählt. 12 der Arbeiten lagen im oberen, 10 im mittleren und 8 im unteren Notendrittel. Bei den ausgewählten Arbeiten handelt es sich jeweils um eine Kunststoffverblendbrücke auf den Zähnen 24-26. Die ausgewählten Arbeiten wurden durch eine Nummerierung von 1 bis 30 anonymisiert.

3.2 Bewerter

Jede Brücke wurde jeweils vom Studierenden selbst, einem Kommilitonen, zwei Studenten des klinischen Studienabschnitts, zwei vorklinischen Kursassistenten sowie zwei klinischen Kursassistenten bewertet, so dass sich insgesamt vier Bewertergruppen ergaben (Tab. 5).

Bewertergruppe	Bewerter	Tätigkeit	Berufserfahrung
Gruppe I	Studenten	Studenten der Vorklinik; Kursteilnehmer, Phantom-Kurs II	klinisch keine, zwei technische Kurse absolviert
Gruppe II	1 und 2	Betreuung der klinischen Behandlungskurse	Examen vor 3 bzw. 7 Jahren; Studentenbetreuung über 7 bzw. 14 Semester (vorklinisch und klinische Kurse)
Gruppe III	3 und 4	Betreuung des technisch propädeutischen Kurses im 1. Semester	Examen vor 8 bzw. 10 Monaten, Studentenbetreuung über 1 ½ bzw. 2 Semester
Gruppe IV	5 und 6	Studenten des klinischen Studienabschnittes; Kursteilnehmer Behandlungskurs II der Prothetik	vorklinische Ausbildung abgeschlossen; drei klinische Behandlungskurse (prothetisch und konservierende Zahnheilkunde) absolviert

Tab. 5 *Bewertergruppen.*

3.3 Vorgehen bei der Bewertung

Die Bewertung der Brücken durch die jeweiligen Bewerter erfolgte dergestalt, dass zunächst anhand einer **Checkliste** die Einzelaspekte Randschluß, Passgenauigkeit, Approximalkontakt, Brückenglied, Okklusion/Höhe, Okklusalflächengestaltung/ Zahnform, technische Verarbeitung und Ästhetik benotet wurden (siehe Abb.1), wobei keine Bewertungskriterien vorgegeben waren. Als Teilnoten kamen „optimal“, „gut“, „noch akzeptabel“ und „nicht akzeptabel“ in Frage. Dabei liegt die Bestehensgrenze zwischen „noch akzeptabel“ und „nicht akzeptabel“. Ausgehend von den Teilnoten vergab jeder Bewerter dann eine Gesamtbeurteilung von 1 (sehr gut) bis 6 (schlecht) für die jeweilige Arbeit. Es wurde vorher festgelegt, dass ein „nicht akzeptabel“ bei den Kriterien Randschluß und Passgenauigkeit sofort zum Nichtbestehen der Arbeit führt und mit der Gesamtnote 4- oder schlechter bewertet werden musste. Die Gewichtung der restlichen Kriterien für die Gesamtnote lag hingegen im Ermessen des Bewerters.

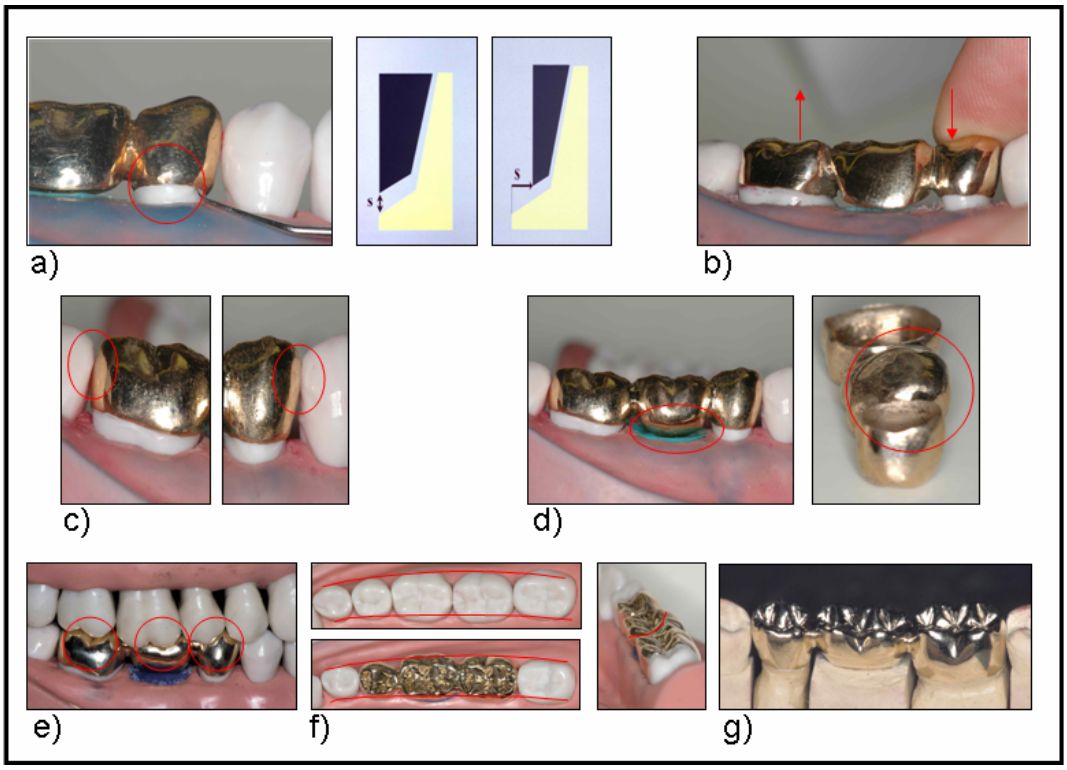
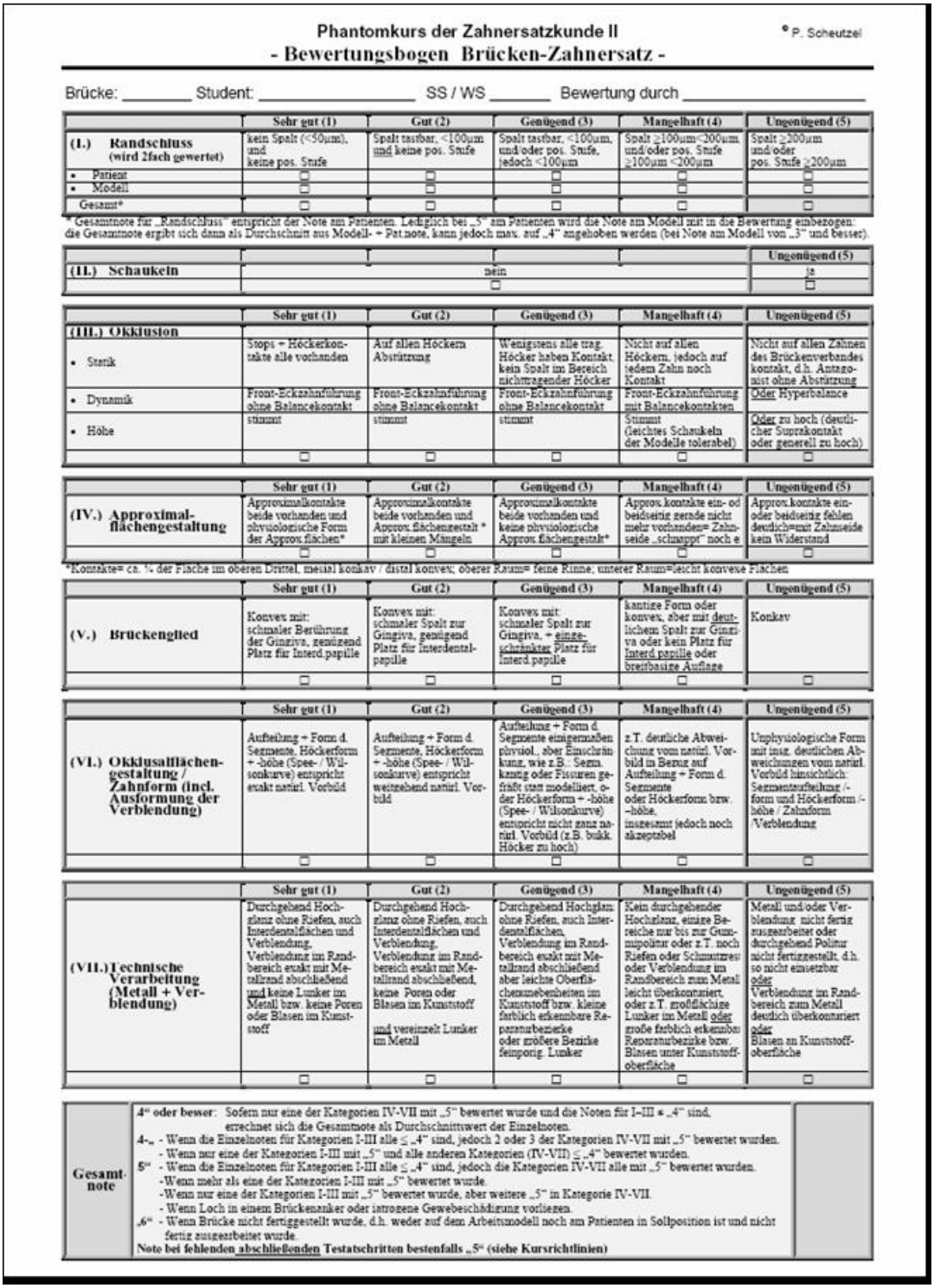


Abb. 1 *Darstellung der bewerteten Teilaspekte: a) Randschluss, b) Passgenauigkeit, c) Approximalkontakt, d) Brückenglied, e) Okklusion/Höhe, f) Okklusalflächengestaltung/Zahnform, g) technische Verarbeitung/Ästhetik*

Zur Feststellung der intraindividuellen Reliabilität wurde diese Bewertung nach einer Woche wiederholt.

Danach erfolgte eine erneute Bewertung mittels eines an der Poliklinik für zahnärztliche Prothetik des Universitätsklinikums Münster entwickelten strukturierten Bewertungsbogens (© P. Scheutzel) (Abb. 2a). Hierbei wurden ebenfalls die Teilaspekte Randschluß, Passgenauigkeit, Approximalkontakt, Brückenglied, Okklusion/Höhe, Okklusalflächengestaltung/Zahnform, technische Verarbeitung und Ästhetik bewertet.



Phantomkurs der Zahnersatzkunde II

© P. Scheutzel

- Bewertungsbogen Brücken-Zahnersatz -

Brücke: _____ Student: _____ SS / WS _____ Bewertung durch _____

	Sehr gut (1)	Gut (2)	Genügend (3)	Mangelhaft (4)	Ungenügend (5)
(I.) **Randschluss** (wird 2fach gewertet)	kein Spalt (<50µm), und keine pos. Stufe	Spalt tastbar, <100µm und keine pos. Stufe	Spalt tastbar, <100µm, und/oder pos. Stufe, jedoch <100µm	Spalt ≥100µm<200µm und/oder pos. Stufe ≥100µm <200µm	Spalt ≥200µm und/oder pos. Stufe ≥200µm
• Patient	☐	☐	☐	☐	☐
• Modell	☐	☐	☐	☐	☐
Gesamt*	☐	☐	☐	☐	☐

* Gesamtnote für „Randschluss" entspricht der Note am Patienten. Lediglich bei „5" am Patienten wird die Note am Modell mit in die Bewertung einbezogen: die Gesamtnote ergibt sich dann als Durchschnitt aus Modell- + Pat.note, kann jedoch max. auf „4" angehoben werden (bei Note am Modell von „3" und besser).

		Ungenügend (5)
(II.) **Schaukeln**	nein	ja
	☐	☐

	Sehr gut (1)	Gut (2)	Genügend (3)	Mangelhaft (4)	Ungenügend (5)
(III.) Okklusion					
• Statik	Stops + Höckerkontakte alle vorhanden	Auf allen Höckern Abstützung	Wenigstens alle trag. Höcker haben Kontakt, kein Spalt im Bereich nichttragender Höcker	Nicht auf allen Höckern, jedoch auf jedem Zahn noch Kontakt	Nicht auf allen Zähnen des Brückenverbandes kontakt, d.h. Antagonist ohne Abstützung
• Dynamik	Front-Eckzahnführung ohne Balancekontakt	Front-Eckzahnführung ohne Balancekontakt	Front-Eckzahnführung ohne Balancekontakt	Front-Eckzahnführung mit Balancekontakten	Oder Hypersbalance
• Höhe	stimmt	stimmt	stimmt	Stimmt (leichtes Schaukeln der Modelle tolerabel)	Oder zu hoch (deutlicher Suprakontakt oder generell zu hoch)
	☐	☐	☐	☐	☐

	Sehr gut (1)	Gut (2)	Genügend (3)	Mangelhaft (4)	Ungenügend (5)
(IV.) **Approximalflächengestaltung**	Approximalkontakte beide vorhanden und physiologische Form der Approx.flächen*	Approximalkontakte beide vorhanden und Approx.flächengestalt * mit kleinen Mängeln	Approximalkontakte beide vorhanden und keine physiologische Approx.flächengestalt*	Approx.kontakte ein- od beidseitig gerade nicht mehr vorhanden= Zahnseide „schnappt" noch e	Approx.kontakte ein- oder beidseitig fehlen deutlich=mit Zahnseide kein Widerstand
	☐	☐	☐	☐	☐

*Kontakte= ca. ¼ der Fläche im oberen Drittel, mesial konkav / distal konvex; oberer Raum= feine Rinne, unterer Raum=leicht konvexe Flächen

	Sehr gut (1)	Gut (2)	Genügend (3)	Mangelhaft (4)	Ungenügend (5)
(V.) **Brückenglied**	Konvex mit: schmaler Berührung der Gingiva, genügend Platz für Interd.papille	Konvex mit: schmaler Spalt zur Gingiva, genügend Platz für Interdentalpapille	Konvex mit: schmaler Spalt zur Gingiva, + eingeschränkter Platz für Interd.papille	kantige Form oder konvex, aber mit deutlichem Spalt zur Gingiva oder kein Platz für Interd.papille oder breitbasige Auflage	Konkav
	☐	☐	☐	☐	☐

	Sehr gut (1)	Gut (2)	Genügend (3)	Mangelhaft (4)	Ungenügend (5)
(VI.) **Okklusalflächengestaltung / Zahnform (incl. Ausformung der Verblendung)**	Aufteilung + Form d. Segmente, Höckerform + -höhe (Spee- / Wilsonkurve) entspricht exakt natürl. Vorbild	Aufteilung + Form d. Segmente, Höckerform + -höhe (Spee- / Wilsonkurve) entspricht weitestgehend natürl. Vorbild	Aufteilung + Form d. Segmente einigermaßen physiol., aber Einschränkung, wie z.B. Segm. kantig oder Fissuren gefräst statt modelliert, oder Höckerform + -höhe (Spee- / Wilsonkurve) entspricht nicht ganz natürl. Vorbild (z.B. bukk. Höcker zu hoch)	z.T. deutliche Abweichung vom natürl. Vorbild in Bezug auf Aufteilung + Form d. Segmente oder Höckerform bzw. -höhe, insgesamt jedoch noch akzeptabel	Unphysiologische Form mit insg. deutlichen Abweichungen vom natürl. Vorbild hinsichtlich Segmentaufteilung / -form und Höckerform / -höhe / Zahnform /Verblendung
	☐	☐	☐	☐	☐

	Sehr gut (1)	Gut (2)	Genügend (3)	Mangelhaft (4)	Ungenügend (5)
(VII.) **Technische Verarbeitung (Metall + Verblendung)**	Durchgehend Hochglanz ohne Riefen, auch Interdentalflächen und Verblendung, Verblendung im Randbereich exakt mit Metallrand abschließend und keine Lunker im Metall bzw. keine Poren oder Blasen im Kunststoff	Durchgehend Hochglanz ohne Riefen, auch Interdentalflächen und Verblendung, Verblendung im Randbereich exakt mit Metallrand abschließend, keine Poren oder Blasen im Kunststoff und vereinzelt Lunker im Metall	Durchgehend Hochglanz ohne Riefen, auch Interdentalflächen, Verblendung im Randbereich exakt mit Metallrand abschließend aber leichte Oberflächenunebenheiten im Kunststoff bzw. kleine farblich erkennbare Reparaturbereiche oder größere Bezirke feinporig. Lunker	Kein durchgehender Hochglanz, einige Bereiche nur bis zur Grundpolitur oder z.T. noch Riefen oder Schmutzreste oder Verblendung im Randbereich zum Metall leicht überkontariert, oder z.T. großflächige Lunker im Metall oder große farblich erkennbare Reparaturbezirke bzw. Blasen unter Kunststoffoberfläche	Metall und/oder Verblendung nicht fertig ausgearbeitet oder durchgehend Politur nicht fertiggestellt, d.h. so nicht einsetzbar oder Verblendung im Randbereich zum Metall deutlich überkonturiert oder Blasen an Kunststoffoberfläche
	☐	☐	☐	☐	☐

Gesamtnote

4" oder besser: Sofern nur eine der Kategorien IV-VII mit „5" bewertet wurde und die Noten für I–III ≤ „4" sind, errechnet sich die Gesamtnote als Durchschnittswert der Einzelnoten.

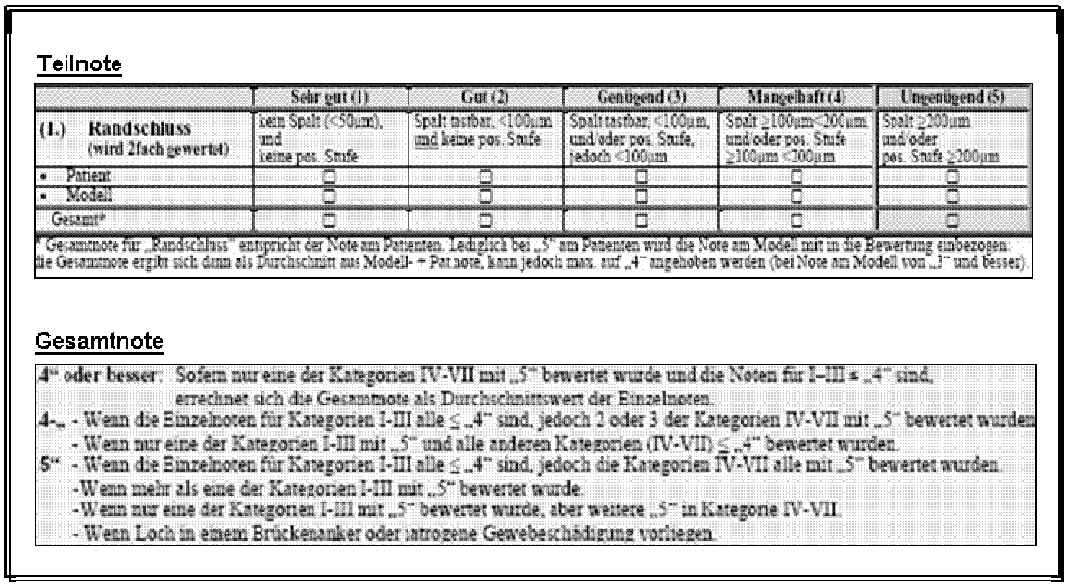
4-„ - Wenn die Einzelnoten für Kategorien I-III alle ≤ „4" sind, jedoch 2 oder 3 der Kategorien IV-VII mit „5" bewertet wurden.
- Wenn nur eine der Kategorien I-III mit „5" und alle anderen Kategorien (IV-VII) ≤ „4" bewertet wurden.

5" - Wenn die Einzelnoten für Kategorien I-III alle ≤ „4" sind, jedoch die Kategorien IV-VII alle mit „5" bewertet wurden.
- Wenn mehr als eine der Kategorien I-III mit „5" bewertet wurde.
- Wenn nur eine der Kategorien I-III mit „5" bewertet wurde, aber weitere „5" in Kategorie IV-VII.
- Wenn Loch in einem Brückenanker oder iatrogene Gewebeschädigung vorliegen.

„6" - Wenn Brücke nicht fertiggestellt wurde, d.h. weder auf dem Arbeitsmodell noch am Patienten in Sollposition ist und nicht fertig ausgearbeitet wurde.

Note bei fehlenden abschließenden Testatschritten bestenfalls „5" (siehe Kursrichtlinien)

Abb. 2a *Bewertungsbogen Brücke-Zahnersatz (Übersicht).*



Teilnote

	Sehr gut (1)	Gut (2)	Genügend (3)	Mangelhaft (4)	Ungenügend (5)
(1.) **Randschluss** (wird 2fach gewertet)	kein Spalt (<50µm), und keine pos. Stufe	Spalt tastbar, <100µm und keine pos. Stufe	Spalt tastbar, <100µm, und/oder pos. Stufe, jedoch <100µm	Spalt ≥100µm<200µm und/oder pos. Stufe ≥100µm <200µm	Spalt ≥200µm und/oder pos. Stufe ≥200µm
• Patient	☐	☐	☐	☐	☐
• Modell	☐	☐	☐	☐	☐
Gesamt*	☐	☐	☐	☐	☐

* Gesamtnote für „Randschluss" entspricht der Note am Patienten. Lediglich bei „5" am Patienten wird die Note am Modell mit in die Bewertung einbezogen: die Gesamtnote ergibt sich dann als Durchschnitt aus Modell- + Patnote, kann jedoch max. auf „4" angehoben werden (bei Note am Modell von „3" und besser).

Gesamtnote

4" oder besser: Sofern nur eine der Kategorien IV-VII mit „5" bewertet wurde und die Noten für I–III ≤ „4" sind, errechnet sich die Gesamtnote als Durchschnittswert der Einzelnoten.

4-„ - Wenn die Einzelnoten für Kategorien I-III alle ≤ „4" sind, jedoch 2 oder 3 der Kategorien IV-VII mit „5" bewertet wurden
- Wenn nur eine der Kategorien I-III mit „5" und alle anderen Kategorien (IV-VII) ≤ „4" bewertet wurden.

5" - Wenn die Einzelnoten für Kategorien I-III alle ≤ „4" sind, jedoch die Kategorien IV-VII alle mit „5" bewertet wurden
- Wenn mehr als eine der Kategorien I-III mit „5" bewertet wurde.
- Wenn nur eine der Kategorien I-III mit „5" bewertet wurde, aber weitere „5" in Kategorie IV-VII.
- Wenn Loch in einem Brückenanker oder iatrogene Gewebeschädigung vorliegen.

Abb.2b *Bewertungsbogen Brücken Zahnersatz (Ausschnitt).*

Anstatt der Beurteilung „optimal", „gut", „noch akzeptabel" und „nicht akzeptabel" (siehe oben) wurden jedoch Noten von 1-6 vergeben, wobei 1 „sehr gut" und 6 „schlecht" bedeutete. Die Notenkriterien, wann welche Note vergeben wird, waren anhand des Bewertungsbogens genau definiert. So wurde zum Beispiel der Randschluss „gut" (2) benotet, wenn der Kronenrandspalt <100µm war. Die Gesamtnote der Brücke wurde abschließend ebenfalls nach genauen Vorgaben des Bewertungsbogens ermittelt (Abb. 2b).

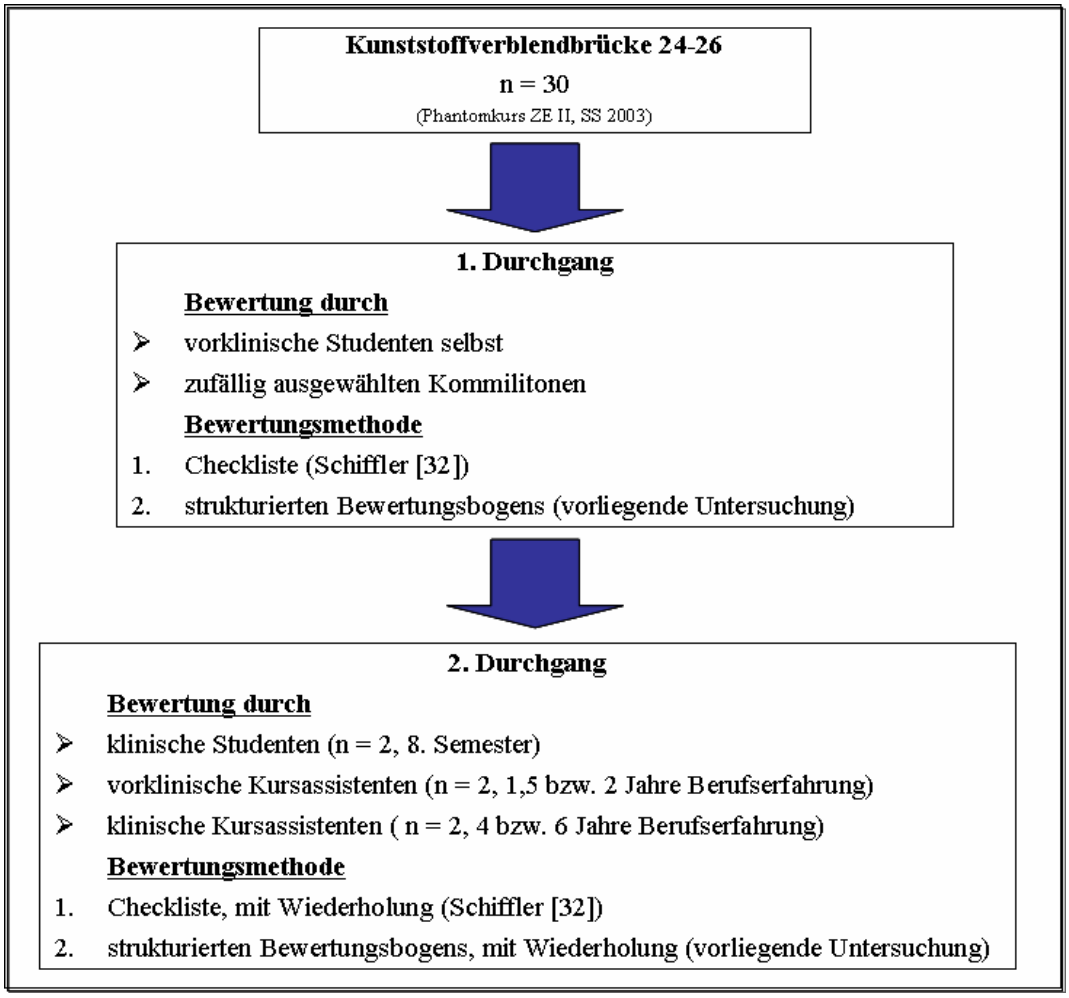


Abb. 3 *Schematische Darstellung des Untersuchungsablaufes.*

Die beiden Untersuchungsreihen liefen dabei wie folgt ab (siehe Abb. 3):

- Bewertung außerhalb des Phantomkopfes unter Gesichtspunkten der Benotung von Phantomarbeiten mit Checkliste – 2 Durchgänge
- Bewertung außerhalb des Phantomkopfes unter Gesichtspunkten der Benotung von Phantomarbeiten unter Zuhilfenahme des Bewertungsbogens – 2 Durchgänge

3.4 Statistische Auswertung

Die Auswertung der Daten erfolgte als explorative Datenanalyse auf einem Notebook mit 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 [30], die Varianzhomogenität mit dem Levene-Test [30] geprüft.

Zur Beurteilung der **Objektivität**, d.h. der **interindividuellen Reliabilität** wurde zunächst 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 [30]. Hierbei wird von einem statistisch signifikanten Ereignis gesprochen, wenn die Irrtumswahrscheinlichkeit für den Fehler 1. Art (a) unter 5% liegt ($p<0,05$).

Als Maß für die **interindividuelle Reliabilität** wurde außerdem der Intraklassenkorrelationskoeffizient (ICC=Intraclass-Correlaion-Coefficient) herangezogen.

Die Intraklassenkorrelation ist ein parametrisches Verfahren zur Quantifizierung der Übereinstimmung zwischen mehreren Beurteilern in Bezug auf mehrere Beobachtungsobjekte. Das Verfahren der Intraklassenkorrelation stellt eine Varianzanalyse dar, mit der das Varianzverhältnis von „wahrer“ Varianz, d.h. der tatsächlich durch die zu bewertende Objekte selber verursachten Varianz, zur Gesamtvarianz ermittelt werden kann. Somit bedeutet ein ICC-Wert von 0,9, dass die Einschätzung eines Beurteilers zu 90% von den tatsächlichen „wahren“ Werten der beurteilten Person determiniert werden und lediglich 10% der Varianz der Daten durch Fehlereinflüsse bestimmt sind.

Insgesamt lassen sich bis zu sechs verschiedene Arten des ICC unterscheiden [33], je nachdem ob alle Rater alle oder verschiedene Fälle einschätzen oder ob die Rater zufällig aus einer größeren Menge von Ratern ausgewählt wurden oder nicht, wobei auch zu unterscheiden ist, ob die absoluten Messwerte verglichen werden oder ob das individuelle Mittelwertsniveau der Beurteiler zur Beurteilung der Reliabilität herangezogen wird. Außerdem macht es einen Unterschied, ob die Einzelwerte der Rater miteinander verglichen werden oder es (z.B. um die Stabilität zu erhöhen) um gemittelte Einschätzungen einer Ratergruppe handelt, was z.B. dann der Fall ist, wenn auch in der tatsächlichen Praxis stets mehrere Bewerter die Beurteilung gemeinsam vornehmen.

Im vorliegenden Fall wurde zur Bestimmung der interpersonellen Urteilskonkordanz innerhalb der verschiedenen Bewertergruppen (d.h. zwischen jeweils zwei Bewertern derselben Gruppe) der ICC nach dem *two-way random effects model* vom *consistency type* für *single rater* berechnet. In dieser justierten Form entspricht der ICC dem Produkt-Moment-Korrelationskoeffizienten nach Pearson (rp), welcher einen Spezialfall der Intraklassenkorrelation für die Beurteilung der Übereinstimmung zweier Bewerter darstellt (*Shrout und Fleiss* [33], *Wirtz und Casper* [41]).

Auf diese Weise konnten die eigenen Ergebnisse unmittelbar mit den von *Schiffler* [32] bei der Beurteilung derselben Studienobjekte nach dem herkömmlichen „glance & grade"-System mittels Checkliste ermittelten Korrelationswerten (rp) für die interpersonelle Urteilskonkordanz verglichen werden.

Als Maß für die **intraindividuelle Übereinstimmung,** d.h. die **Reproduzierbarkeit** der Bewertungen jedes einzelnen Bewerters bei Wiederholung der Bewertung wurde der Produkt-Moment-Korrelationskoeffizient nach Pearson (r_p) bestimmt [30]. 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.

4 Ergebnisse

4.1 Vergleich der Durchschnittsnoten verschiedener Bewertergruppen bei der Benotung mittels Bewertungsbogen

Die Durchschnittsnoten der verschiedenen Bewertergruppen sind in der Abb. 4 dargestellt. Es zeigt sich, dass zwischen den Bewertergruppen vorklinische Zahnärzte, klinische Zahnärzte und klinische Studenten kein statistisch signifikanter Unterschied in der Durchschnittsnote bestand. Die Durchschnittsnoten lagen zwischen 3,69 und 3,61 ($p = 0,421$). Lediglich die von den vorklinischen Studenten vergebenen Durchschnittsnoten unterschieden sich von denjenigen aller anderen Bewertergruppen ($p < 0,001$). So lag die Durchschnittnote bei der Bewertung durch die vorklinischen Studenten bei 2,86 und damit im Vergleich zu den Durchschnittsnoten der Bewertergruppen der vorklinischen Zahnärzte, klinischen Zahnärzte und klinischen Studenten (3,61 – 3,69) deutlich niedriger.

In der Aufschlüsselung der der einzelnen Notenbereiche von Kursnote <3,0, Kursnote ≥3,0 über Kursnote >4,0 spiegelt sich ein ähnliches Bild wider.
Betrachtet man die Notenaufschlüsselung für einzelne Teilbereiche der Brücke in Abb. 5, zeigt sich kein einheitliches Bild in der Notenvergabe. Lediglich die Tendenz des Gesamtnotendurchschnitts ist wieder zu erkennen. Beim Kriterium „Paßgenauigkeit“ muss beachtet werden, dass der große Notenunterschied zwischen den vorklinischen Studenten und den restlichen Bewertergruppen dadurch zustande kommt, dass hier nur mit den beiden Noten 1 (bestanden) bzw. 5 (nicht bestanden) bewertet werden konnte, also nur zwei und nicht sechs Noten zur Auswahl standen, also ein relativ großer numerischer Abstand zwischen den zu wählenden Noten liegt, so dass das Kriterium „Paßgenauigkeit“ nur einzeln für sich, oder im Bezug auf die Reliabilität zu beurteilen ist.

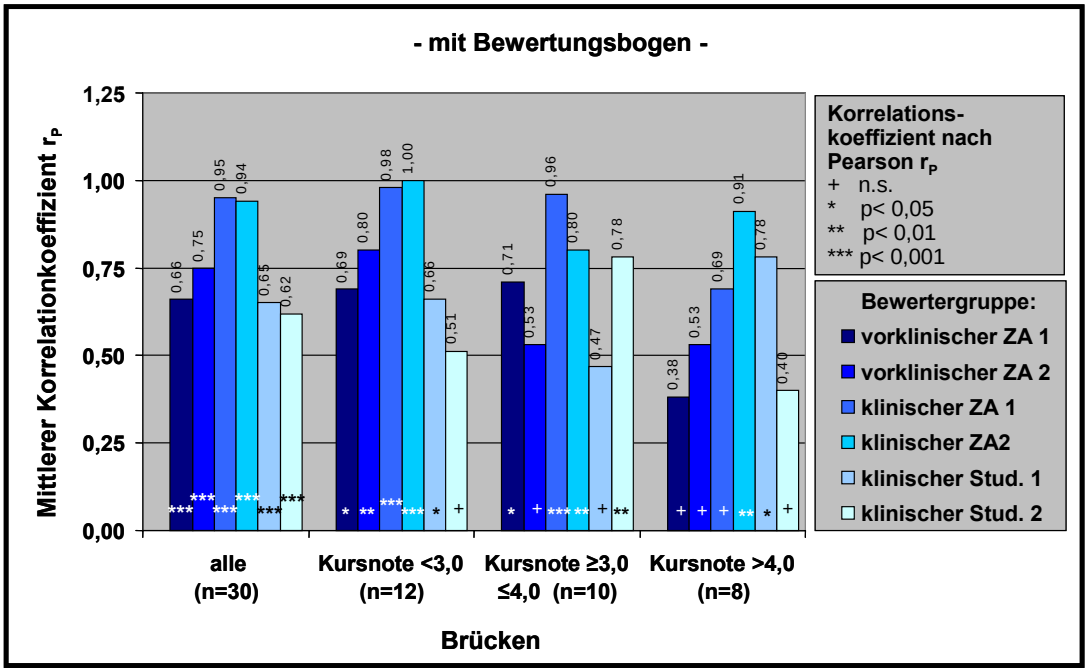


Abb. 4 *Mittlere Durchschnittsnoten der verschiedenen Bewertergruppen bei der Benotung mittels Bewertungsbogen für alle untersuchten Brücken.*

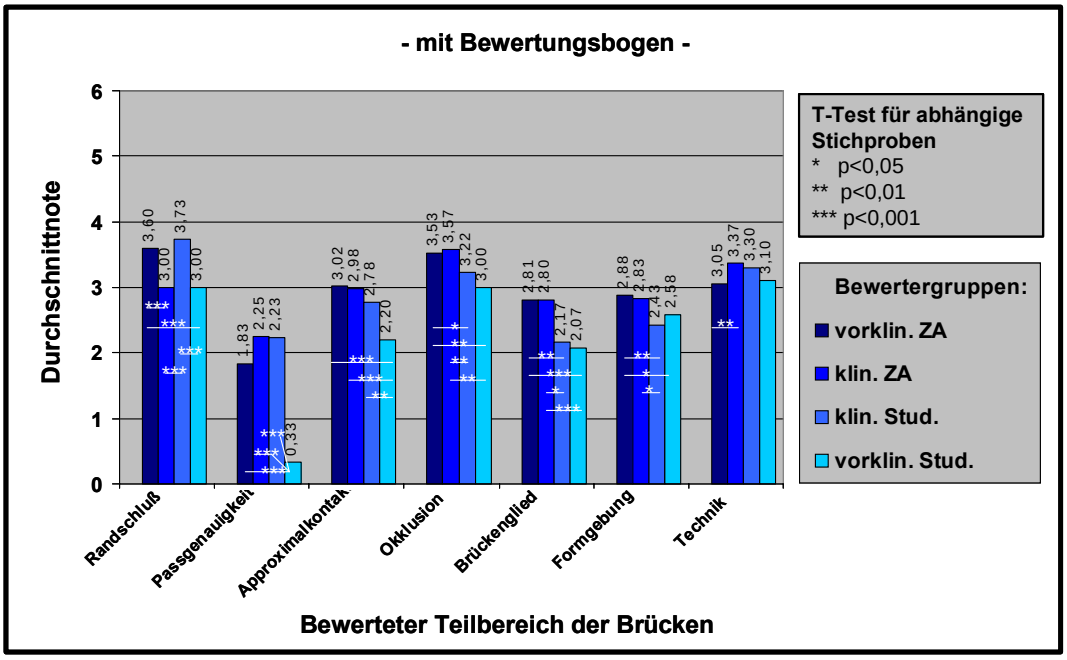


Abb. 5 *Mittlere Durchschnittsnoten der verschiedenen Bewertergruppen bei der Benotung mittels Bewertungsbogen für die einzelnen Teilbereiche der bewerteten Brücken.*

4.2 Vergleich der Notendifferenzen innerhalb der jeweiligen Bewertergruppen bei der Benotung mittels Bewertungsbogen

Die Abb. 6 zeigt die Notendifferenz innerhalb der Bewertergruppen. Die durchschnittliche Notendifferenz betrug 0,57, wobei die stärkste Abweichung (0,89) innerhalb der Gruppe der vorklinischen Zahnärzte und die geringste Differenz (0,22) innerhalb der Gruppe der vorklinisch Studierenden zu beobachten war. Die klinischen Zahnärzte und Studenten liegen mit 0,62 bzw. 0,56 relativ nahe am Gesamtdurchschnitt.

Die Tendenz des Gesamtdurchschnitts, dass die Zahnärzte die größte Differenz und die vorklinischen Studenten die geringsten Abweichungen aufweisen, spiegelt sich auch in der Einzelnotenaufschlüsselung wieder. Lediglich im Bereich der Kursnote >4,0 liegt der Durchschnitt bei den vorklinischen Studenten bei 0,76 und weicht damit statistisch signifikant von den Differenzen dieser Bewertergruppe in den anderen Notenbereichen ab.

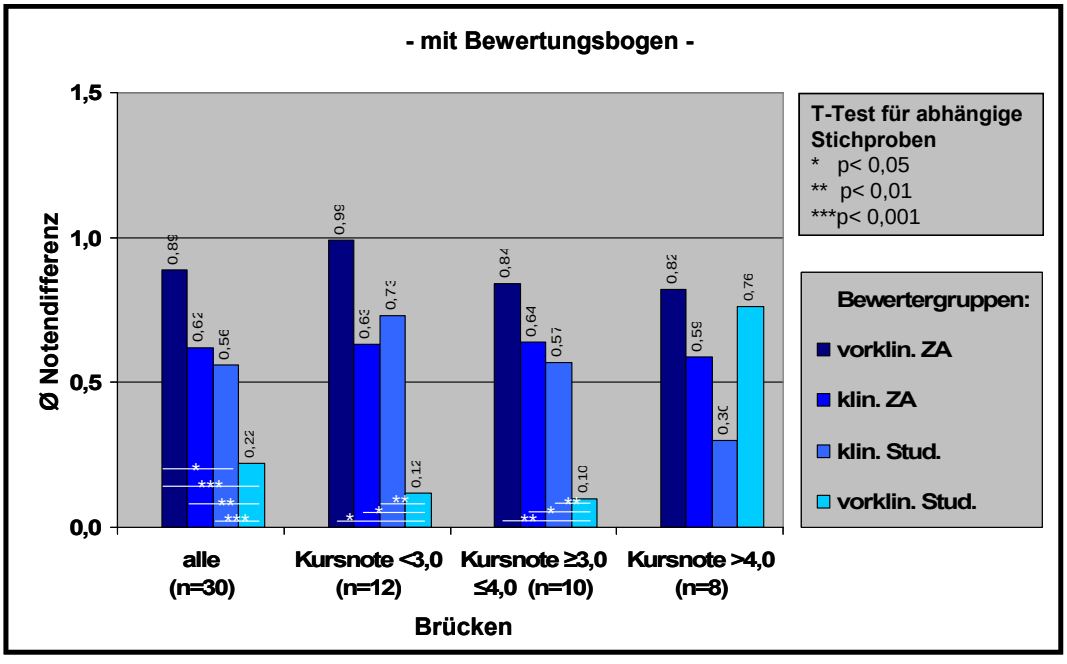


Abb. 6 *Durchschnittliche Notendifferenz bei der Benotung mittels Bewertungsbogen innerhalb der verschiedenen Bewertergruppen.*

Die Aufschlüsselung der Differenzen innerhalb der einzelnen Kriterien zeigt die Abb. 7. Im Gegensatz zu den vorherigen Ergebnissen ergibt sich hierbei innerhalb der verschie-

denen Bewertergruppen kein einheitliches Bild. Die statistisch signifikanten Unterschiede zwischen den Bewertergruppen beim Kriterium „Paßgenauigkeit“ lassen sich wiederum damit begründen, dass es hier nur die Note 1 bzw. 5 zu vergeben gab. Aufgrund dieser relativ großen numerischen Differenz, kommt es zu den dargestellten Differenzen bei den vorklinischen Zahnärzten von durchschnittlich 1,33 Notenpunkten. Auffällig ist, dass es bei den Kriterien, die mit Messwerten am detailiertesten aufgeschlüsselt wurden, wie dem „Randschluß“, „Approximalkontakt“ oder dem „Brückenglied“, zu den größten Differenzen zwischen den Gruppen kam. Bei der „Okklusion“ oder der „Technik“, die durch deskriptive Kriterien bewertet wurden, lagen die geringsten Notendifferenzen vor.

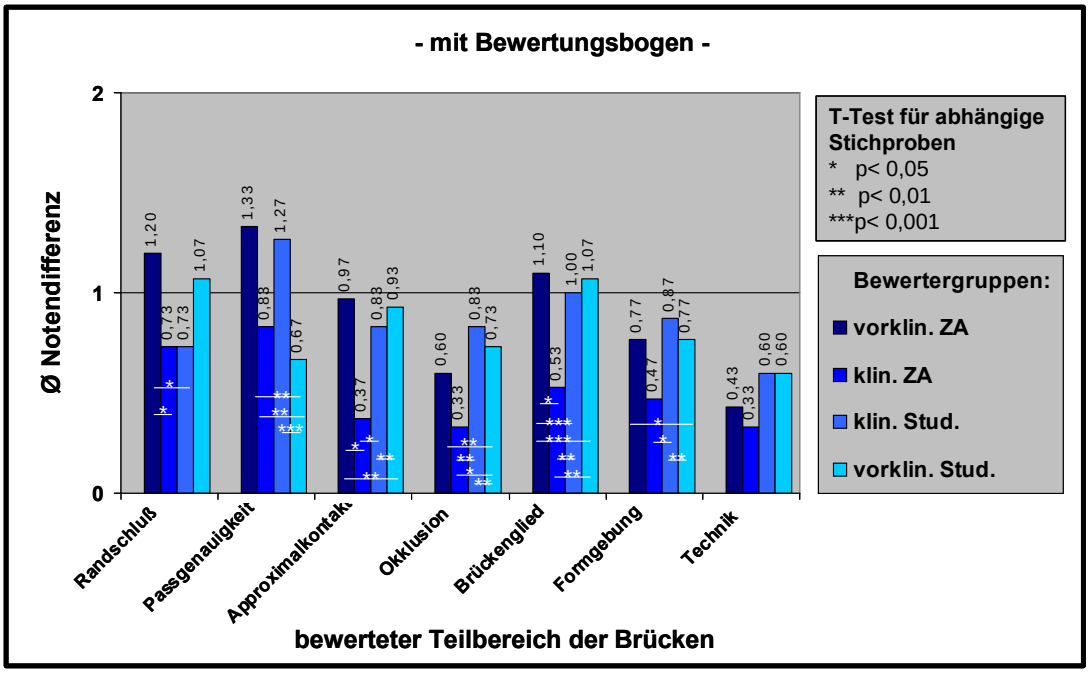


Abb. 7 *Mittlere Durchschnittsnoten der verschiedenen Bewertergruppen bei der Benotung mittels Bewertungsbogen für die einzelnen Teilbereiche der bewerteten Brücken.*

4.3 Vergleich der studentischen Selbsteinschätzung mit dem Urteil anderer Bewerter bei der Benotung mittels Bewertungsbogen

Die Abb. 8 zeigt, dass alle Bewertergruppen mit einer durchschnittlichen Notendifferenz von 0,78 von der studentischen Selbsteinschätzung abwichen und zwar Sinne, dass die Note der Selbsteinschätzung durchweg besser ausfiel als die der Fremdbewertung. Die Gruppenergebnisse liegen mit einer durchschnittlichen Notendifferenz von 0,74 für die klinischen Zahnärzte und 0,83 für die vorklinischen Zahnärzte sehr nah beieinander. Dieser geringe Unterschied spricht für eine gute Kalibrierung der Bewertergruppen durch den Bewertungsbogen.

Auffällig ist, dass die größte Abweichung im oberen (0,92) und die geringste im unteren Notendrittel (0,68) liegt. Wie die Abb. 8 zeigt, ist dies damit zu erklären, dass die vorklinischen Studenten im Bereich der Kursnote < 3,0 tendenziell besser bewertet haben, und so die größere Notendifferenz zu den anderen Bewertergruppen zustande kommt.

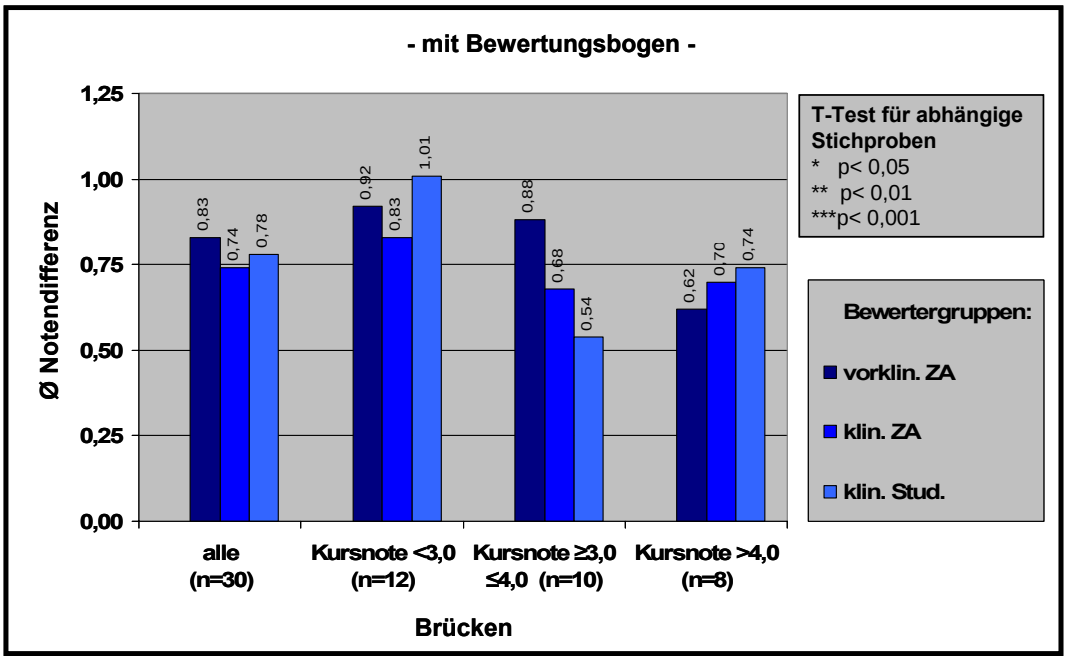


Abb. 8 *Durchschnittliche Abweichung der Selbstbenotung durch vorklinischen Studenten von der Fremdbewertung durch andere Bewerter bei der Benotung mittels Bewertungsbogen (Differenz = Note (Fremdeinschätzung) – Note (Selbsteinschätzung)).*

4.4 Interpersonelle Urteilskonkordanz innerhalb verschiedener Bewertergruppen bei der Benotung mittels Bewertungsbogen

Misst man das Ausmaß der interindividuellen Urteilsübereinstimmung innerhalb der jeweiligen Bewertergruppen mittels des Intraklassenkorrelationskoefficienten (ICC), so ergaben sich Werte zwischen 0,47 (vorklinische Studenten) und 0,69 (klinische Zahnärzte) (Abb. 9) und damit eine mäßig gute Übereinstimmung. Die jeweils höchste Urteilskonkordanz lag in allen Bewertergruppen bei der Beurteilung der Brücken des mittleren Notensegmentes vor.

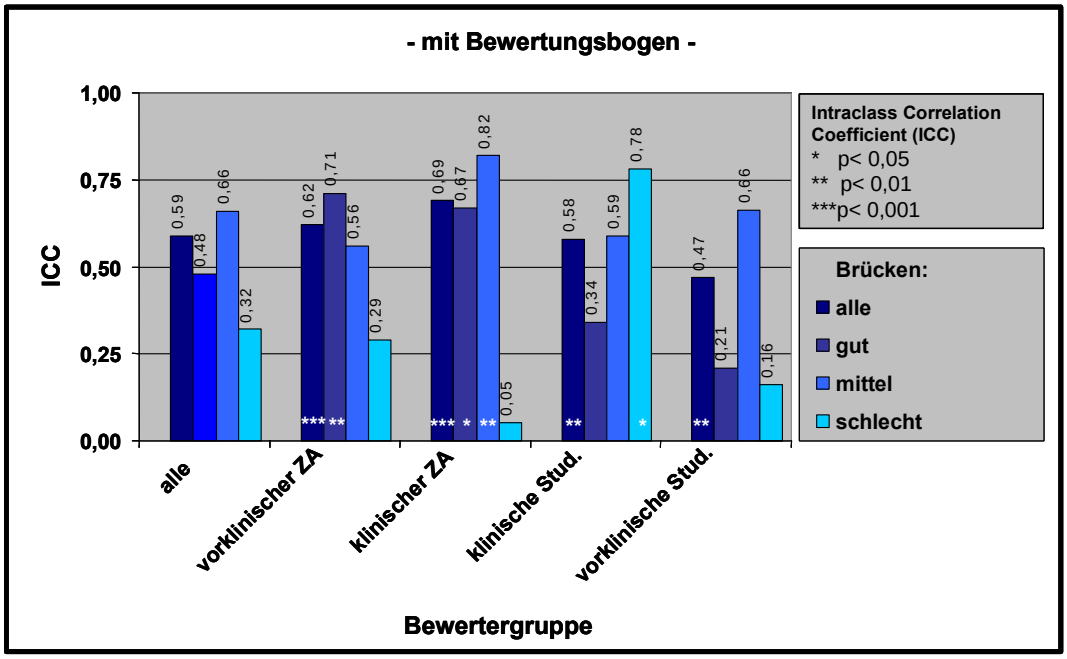


Abb.9 *Mittlerer Intraclass Correlation Coefficient (ICC) als Maß für die interpersonelle Urteilskonkordanz innerhalb verschiedener Bewertergruppen bei der Benotung mittels Bewertungsbogen.*

Abb. 10 zeigt die Aufschlüsselung der einzelnen Teilbereiche für das Ausmaß der interindividuellen Urteilskonkordanz anhand des Intraklassenkorrelationskoeffizienten (ICC) mit Werten zwischen 0,35 (Okklusion) und 0,73 (Technik). Innerhalb der jeweiligen Bewertergruppen zeigen die klinischen Zahnärzte die kleinste Streuung

(0,51-0,86) und die vorklinischen Studenten die größte Streuung (0,1-1,0) für die Teilbereiche. Im Mittel ergibt sich also auch hier eine mäßig gute Übereinstimmung.

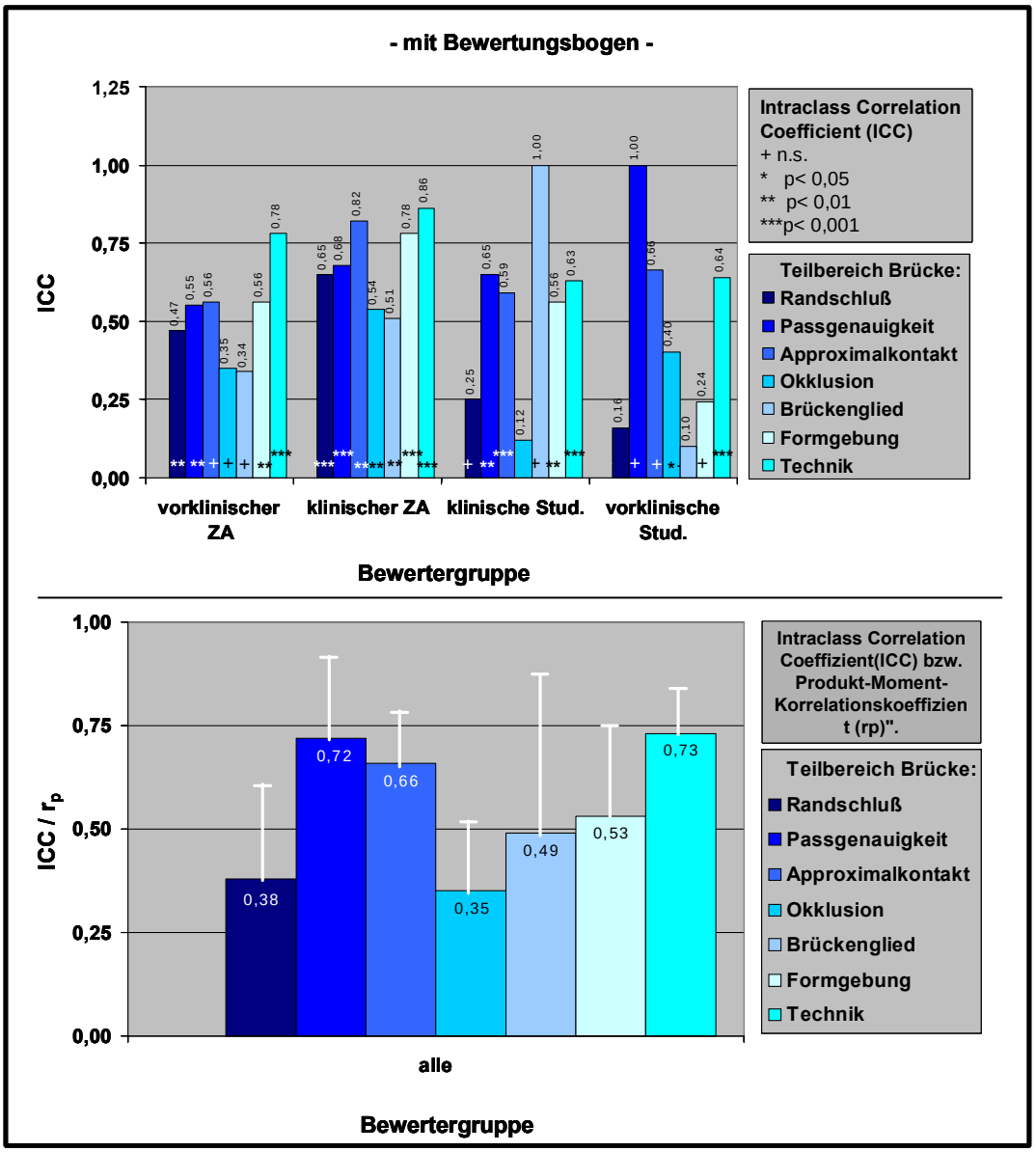


Abb. 10 *Mittlerer Intraclass Correlation Coefficient (ICC) als Maß für die interpersonelle Urteilskonkordanz für die einzelnen Teilbereiche der bewerteten Brücken bei Benotung verschiedener Bewertergruppen.*

4.5 Intrapersonelle Urteilskonkordanz bei der Benotung mittels Bewertungsbogen

Als Maß für diese intraindividuelle Urteilsübereinstimmung wurde der Korrelationskoeffizient nach Pearson bestimmt, der für die Gruppe der klinischen Zahnärzte mit Werten zwischen 0,94 und 0,96 eine hervorragende Reproduzierbarkeit der Benotung ergab. Die niedrigste intrapersonelle Urteilskonkordanz mit r_p-Werten von 0,65 beziehungsweise 0,62 war bei den klinischen Studenten feststellbar. Insgesamt war die Reproduzierbarkeit der intraindividuellen Bewertung mit einem durchschnittlichen Korrelationskoeffizienten von $r_p = 0,76$ sehr hoch.

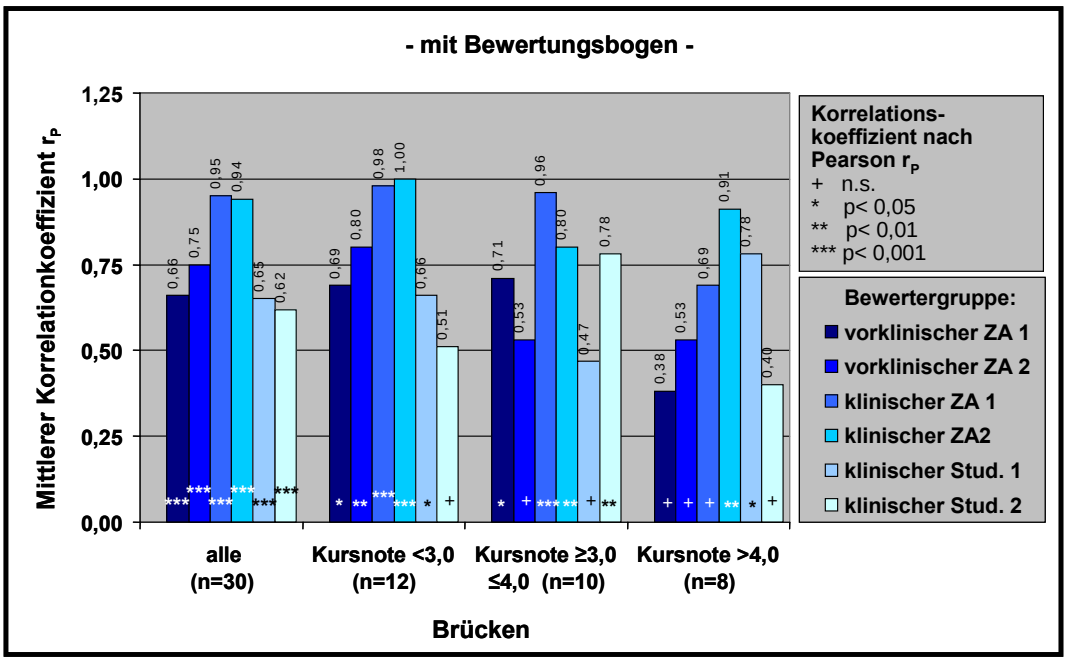


Abb. 11 *Intrapersonelle Urteilskonkordanz bei wiederholter (2-maliger) Bewertung für alle Bewerter bei der Benotung mittels Bewertungsbogen.*

4.6 Vergleich der Bewertungen mit und ohne Bewertungsbogen

Lässt man die Gruppe der vorklinischen Studenten unberücksichtigt, zeigt die Graphik, dass die Bewertung mittels des vorgegebenen Bewertungsbogens zu einer signifikanten

Übereinstimmung geführt hat. So ergibt sich eine Differenz von 0,08 Notenpunkten mit Bewertungsbogen im Vergleich zu 0,60 Notenpunkten ohne Bewertungsbogen [32] zwischen den Bewertergruppen II bis IV (Abb. 12).

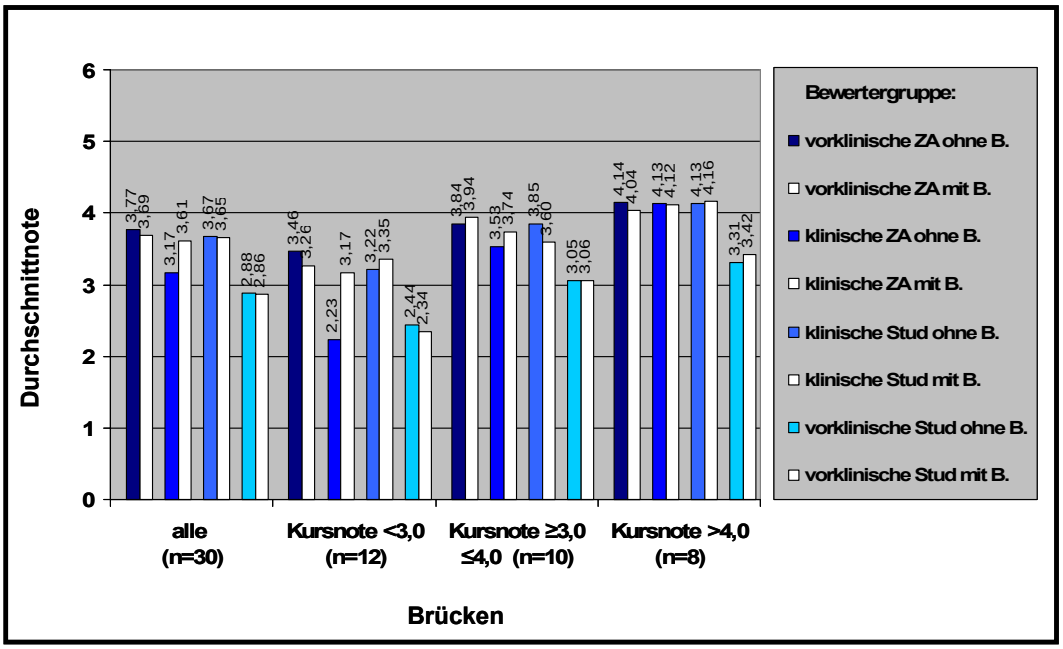


Abb. 12 *Vergleich der Bewertungen mit und ohne Bewertungsbogen (Werte ohne Bewertungsbogen aus [32]).*

Was die Notendifferenz innerhalb der jeweiligen Bewertergruppe und damit die interindividuelle Übereinstimmung betrifft, so ergab sich durch den Einsatz des Bewertungsbogens insgesamt eine geringfügige Verbesserung, das heißt eine Verringerung der Notendifferenz (Abb. 13). Die am stärksten in der Gruppe der vorklinischen Zahnärzte ausgeprägt war. Dennoch spiegelt sich hier die gleiche Verbesserungstendenz wie bei der Notenübersicht wider, denn auch hier fällt die Benotung mit Bogen besser aus. Die Differenz lag im Durchschnitt bei 0,69 Notenpunkten mit Bogen und bei bis zu 0,77 Notenpunkten ohne Bogen [32].

Am Auffälligsten ist der Unterschied zwischen den Bewertungsergebnissen mit und ohne Bogen bei der Gruppe der vorklinischen Zahnärzte. Im Durchschnitt sank hier die Differenz von 1,13 auf 0,89 Notenpunkte.

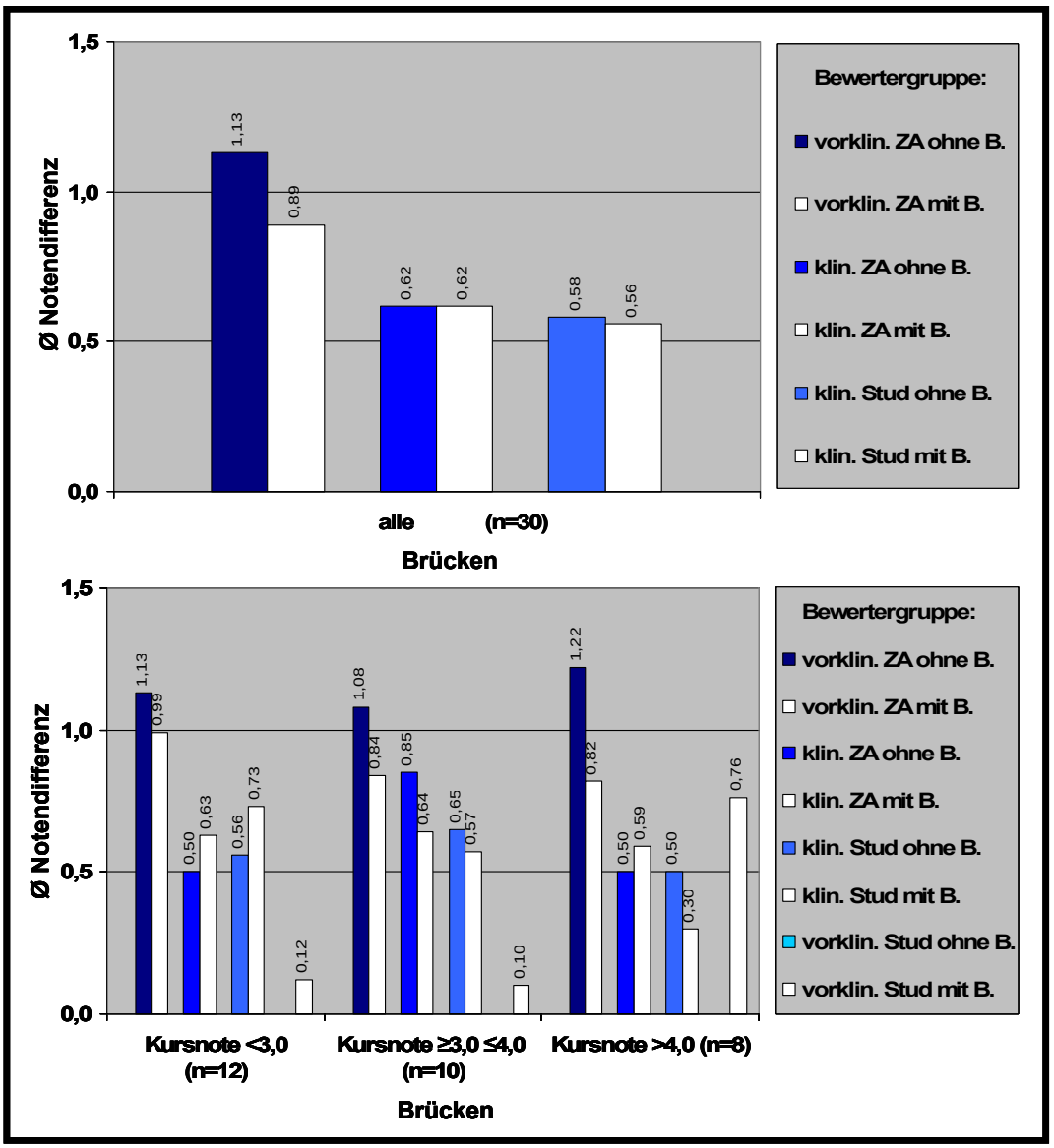


Abb. 13 *Einfluss des Bewertungsbogens auf die Notendifferenz (Werte ohne Bewertungsbogen aus [32]).*

Die Abb. 14 zeigt den Vergleich der einzelnen Bewertergruppen bezüglich der interpersonellen Urteilskonkordanz mit und ohne Bewertungsbogen.

Eine Verbesserung der Urteilskonkordanz durch den Bewertungsbogen ist hier vor allem innerhalb der Gruppe der vorklinischen Zahnärzte zu erkennen, bei denen der ICC durch den Einsatz des Bewertungsbogens eine eindeutige Steigerung von 0,11 „ohne Bewertungsbogen“ [32] auf 0,62 „mit Bewertungsbogen“ anstieg. Bei den klinischen Studenten führte der Einsatz des Bogens dagegen zu einer Absenkung des Koeffizienten und damit zu einem Verlust an Reliabilität. Am Auffälligsten ist der Einbruch bei den klinischen Zahnärzten im Bereich der schlechten Brücken. Hier ging die sehr gute Urteilskonkordanz von 0,94 „ohne Bewertungsbogen“ [32] auf einen Korrelationskoeffizienten von 0,05 „mit Bewertungsbogen“ zurück.

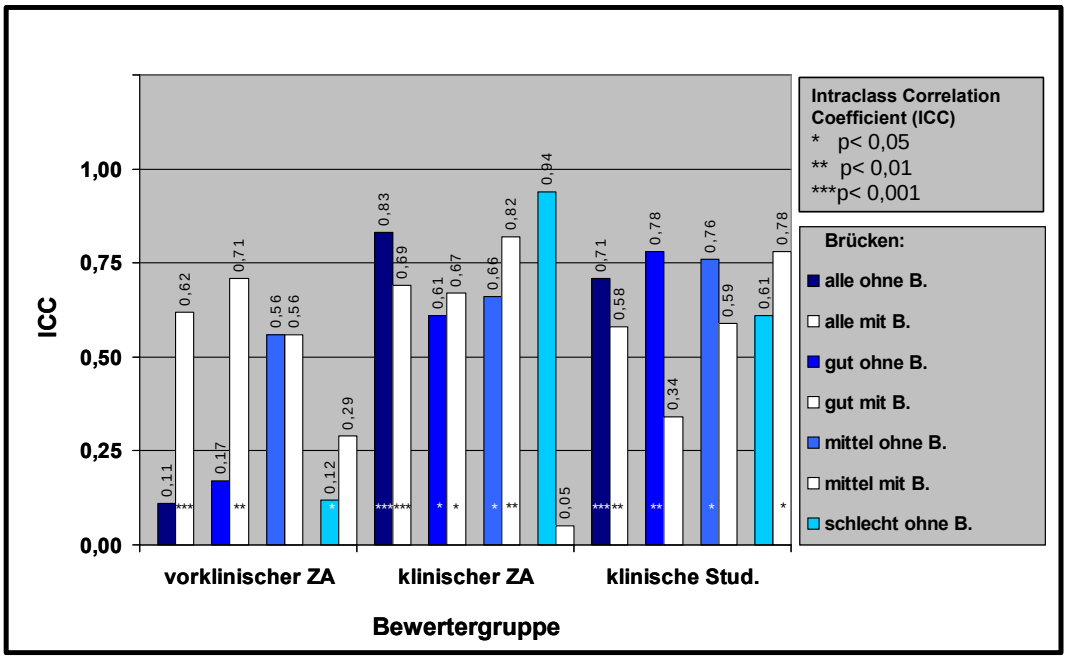


Abb. 14 *Einfluss des Bewertungsbogens auf die interpersonelle Unteilskonkordanz (Werte ohne Bewertungsbogen aus [32]).*

Was die intraindividuelle Urteilsübereinstimmung betrifft, so konnten alle Bewerter bis auf die beiden klinischen Studenten durch den Einsatz des Bewertungsbogens ihre individuelle Urteilskonkordanz deutlich steigern (Abb. 15).

Die größte Steigerung zeigte sich bei dem vorklinischen Zahnarzt 1 von einem Korrelationskoeffizienten von 0,08 ohne definierte Bewertungsbogen [32] auf 0,66 mit definierten Bewertungsbogen. Selbst die hervorragende Übereinstimmung der Gruppe der klinischen Zahnärzte bei der Bewertung ohne den definierten Bewertungsbogen mit

Werten von 0,72 und 0,83 [32], konnte mittels des definierten Bewertungsbogens noch mal auf Werte zwischen 0,94 und 0,95 gesteigert werden.

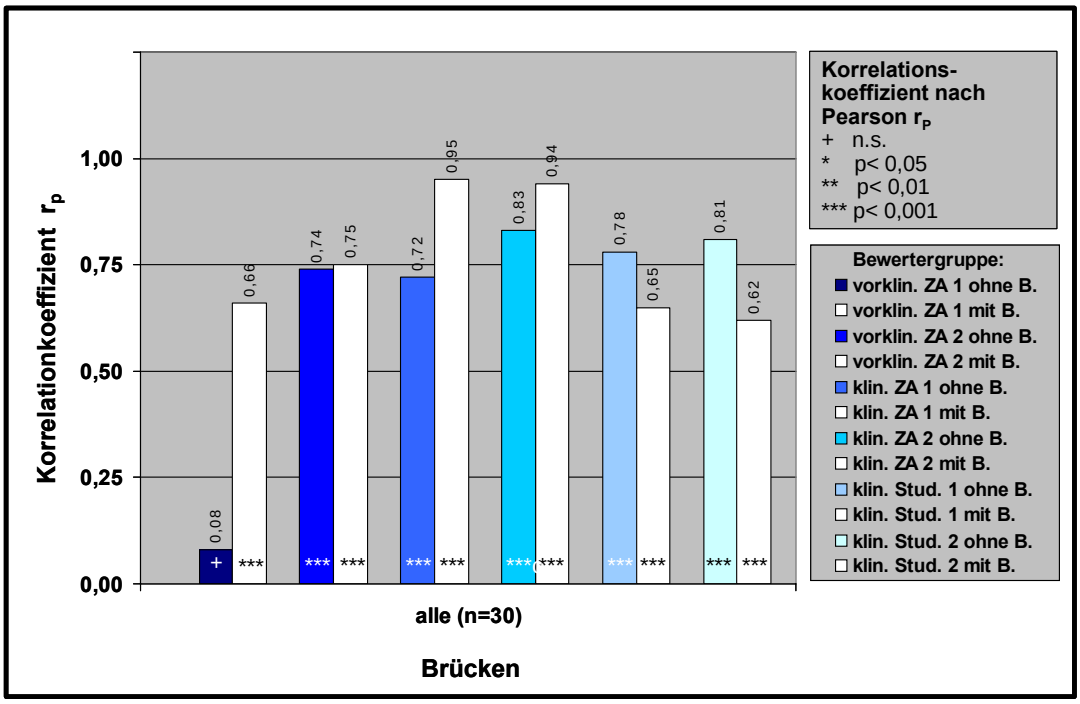


Abb. 15 *Einfluss des Bewertungsbogens auf die intrapersonelle Urteilskonkordanz (Korrelationskoeffizient nach Pearson, r_p) (Werte ohne Bewertungsbogen aus [32]).*

5 Diskussion

In der Medizin allgemein und speziell in der klinischen Zahnmedizin spielt die Reliabilität eine große Rolle [8, 1, 27, 37, 38, 42, 43]. Insbesondere bei den diagnostischen Fragestellungen wird der Reliabilität bei der Leistungsmessung und –beurteilung in den universitären Kursen eine tragende Bedeutung beigemessen.
Die allgemein bekannte Problematik, dass es bei nahezu gleichen Bewertungsbedingungen zum Teil zu erheblich Unterschieden in der Benotungen gleicher Arbeiten durch verschiedene Prüfer kommt, ist vor allem in den praktischen Kursen und hier besonders in den vorklinischen Phantomkursen relevant. Dieses Problem wurde zum einen durch die Arbeiten anderer Autoren [5, 8, 17, 21, 31] als auch durch die Ergebnisse der vorliegenden eigenen Untersuchung bestätigt. Ähnliche Phänomene sind von mündlichen Prüfungen in der Psychologie [2] und in der Medizin [10, 39] seit langem bekannt. Dieser Problematik sollte daher insbesondere bei der Bewertung zahnärztlicher bzw. zahntechnischer Arbeiten eine hohe Aufmerksamkeit gewidmet werden, um zukünftig eine objektiv gerechte Bewertungsgrundlage schaffen zu können.
Die Problematik der geringen Reliabilität und Objektivität in der Beurteilung von vorklinischen zahnärztlichen Arbeiten ist zwar schon lange bekannt, dennoch gibt es bisher nur sehr wenigen Studien, die sich mit diesem Thema beschäftigt haben. Insbesondere in der deutschsprachigen zahnärztlichen Literatur gibt es bisher nur eine Veröffentlichung zum Thema Reliabilität bei der Benotung studentischer Arbeiten [34]. Im Gegensatz dazu spielt die Leistungsmessung und –beurteilung in der pädagogischen Fachliteratur eine weit aus größere Rolle [4, 15, 16, 22, 44].

Diejenigen Untersuchungen, die sich bisher mit dem Thema Reliabilität bei der Benotung zahnmedizinisch studentischer Arbeiten beschäftigt haben, fanden unabhängig der jeweiligen Bewertungsmethode durchweg nur eine geringe inter- und intraindividuelle Reliabilität bei der Bewertung von studentischen Arbeiten. Diese lag in bisherigen Studien zwischen ICC=0,23 und ICC=0,68 bzw. r_s=0,2 und r_s=0,83. Somit ergibt sich bei der Bewertung von vorklinischen Phantomkursarbeiten bisher eine

durchschnittliche **inter**individuelle Reliabilität von ICC = 0,4 [6] bzw. eine **intra**individuelle Reliabilität von r_s / r_p = 0,5.

Vor diesem Hintergrund war es das Ziel der vorliegenden Arbeit festzustellen, inwieweit ein standardisierter Bewertungsbogen mit definierten Kriterien die Reliabilität und Objektivität bei der Bewertung vorklinischer Phantomarbeiten steigern kann.

Die Erwartungen an die vorliegende Studie, eine signifikante Steigerung der Reliabilität und Objektivität durch den Einsatz des strukturierten Bewertungsbogens zu erlangen, konnten allerdings nur teilweise erfüllt werden.

Bezüglich der **Reproduzierbarkeit** (intrapersonelle Urteilskonkordanz) der Benotungen zeigte sich insgesamt eine statistisch signifikante Steigerung der Übereinstimmung von r_p = 0,66 „ohne Bewertungsbogen“ [32] auf r_p = 0,76 „mit Bewertungsbogen“. Lässt man beide Studentengruppen unberücksichtigt und zieht nur die beiden Gruppen der Zahnärzte in die Bewertung mit ein, so wird der Unterschied zwischen den Bewertungen mit und ohne Bewertungsbogen noch deutlicher. In diesem Fall wird eine durchschnittliche Steigerung der **intra**individuellen Reliabilität von r_p = 0,59 „ohne Bewertungsbogen“ [32] auf r_p = 0,83 „mit Bewertungsbogen“ erzielt. Auch wenn man einen gewissen Erinnerungseffekt bei den Wiederholungsdurchgängen mit berücksichtigen muss, wodurch der Korrelationskoeffizient (r_p) gesteigert wird, kann man hier durchaus von einer deutlichen Verbesserung der Zuverlässigkeit bei der Benotung durch den Einsatz des Bewertungsbogens sprechen.

Bezüglich der **Objektivität** (interindividuelle Reliabilität) der Benotung ergab sich zwar keine statistisch signifikante Steigerung im Vergleich der Bewertungen ohne und mit Bewertungsbogen, allerdings nahm die Differenz bei der Benotung innerhalb der Bewertergruppen deutlich ab und lag bei der Bewertung „mit Bewertungsbogen“ nur um 0,08 Notenpunkte auseinander, wogegen durchschnittlich die Notendifferenz bei der Bewertung „ohne Bewertungsbogen“ 0,60 Notenpunkte betrug [32].

Ursache für den Notenunterschied zwischen der Gruppe der vorklinischen Studenten und den anderen Gruppen könnte sein, dass mehrere vorklininische Studenten aus Kollegialität zu ihren Kommilitonen Arbeiten, die im unteren Grenzbereich lagen, tendenziell besser bewertet haben, als sie objektiv tatsächlich waren. Dies führte zur

erhöhten Differenz zu denjenigen Bewertergruppen, die sich objektiv an die Bewertungskriterien gehalten haben.
Lässt man dementsprechend die Gruppe der vorklinischen Studenten außer Acht, so wird ein statistisch signifikanter Unterschied bezüglich der Reliabilität bei der Bewertung „mit“ und „ohne Bewertungsbogen“ deutlich.

Diese Ergebnisse bestätigen die bereits von Dhuru et al., Goepferd et al., Hinkelmann et al., Robertello et al. und Gaines et al. gemachten Beobachtungen, wonach durch Verwendung eines definierten Bewertungsbogens die Reliabilität gesteigert werden kann. So stieg z.B. bei der Untersuchung von Gaines et al. [8] der ICC von 0,26 (mit Checkliste) auf 0,56 (mit Bewertungsbogen) an.
Im Gegensatz dazu stehen allerdings die Ergebnisse von *Vann et al.* [36], die keine Steigerung der Reliabilität durch den Einsatz eines strukturierten Bewertungsbogens in ihren Untersuchungen feststellen konnten. Auch *Fuller et al.* [7] fanden keine signifikante Steigerung der Reliabilität durch den Einsatz eines strukturierten Bewertungsbogens, obwohl der Intraklassenkorrelationskoeffizienten (ICC) von 0,4 auf 0,58 anstieg.

Somit lässt sich zusammenfassend feststellen, dass die Zuhilfenahme eines standardisierten Bewertungsbogens mit definierten Kriterien zwar durchweg zu einer Verbesserung der Reliabilität führt, diese jedoch nicht immer statistisch signifikant ist. Insofern stellt der Einsatz eines Bewertungsbogens bei der Bewertung studentischer Arbeiten eine Verbesserung der Zuverlässigkeit der Bewertung dar. Insbesondere im Vergleich zum „glance and grade“-System bestätigen die Untersuchungsergebnis die Grundhypothese, dass durch die Bewertung mit einem standardisierten Bewertungsbogen eine nachweisbare Verbesserung erzielt werden kann.
Hierbei ist allerdings noch zu beachten, um welche Art von Bewertungskriterien es sich handelt. So kam es in der vorliegenden Untersuchung, insbesondere bei denjenigen Kriterien, die mit Messwerten am detailiertesten aufgeschlüsselt wurden, wie dem „Randschluß“ oder dem „Approximalkontakt“ zu der größten Beuteilungsvariabilität. Bei der „Okklusion“ oder der „Technik“, die durch deskriptive Kriterien bewertet wurden, lagen hingegen die geringsten Notendifferenzen vor. Diese Ergebnisse können

zum Teil damit begründet werden, dass je enger der vorgegebenen Bewertungsrahmen für das Kriterium ist, umso eher Abweichungen in der Messgenauigkeit und Einschätzung der einzelnen Bewerter auftreten können, die wiederum zu einer Senkung der Reliabilität führt, wie von *Houpt und Kress* [14] anhand eines Zwei-Punkte-Bewertungsbogens bereits nachgewiesen wurde.

Hieraus ergibt sich für die Weiterentwicklung des vorliegenden Bewertungsbogens die Empfehlung zur Einschränkung der Notenskala von den bisher benutzten 5 Abstufungen („sehr gut“, „gut“, „genügend“, „mangelhaft“, „ungenügend“) auf nur 3 Abstufungen („optimal oder gut“, akzeptabel“, „nicht akzeptabel“).

6 Zusammenfassung

Das Ziel vorklinischer zahnmedizinischer Behandlungskurse am Phantompatienten ist es, den Studierenden die für die Behandlung „echter“ Patienten im klinischen Studienabschnitt erforderliche kognitive und psychomotorische Kompetenz zu vermitteln. In diesem Zusammenhang müssen die Studierenden z.B. verschiedene Arten von Zahnersatz für ihren Phantompatienten anfertigen und eingliedern, wobei letztendlich die Ergebnisqualität benotet wird. Hierbei stellt sich naturgemäß die Frage nach der Verlässlichkeit der Benotung im Hinblick auf die Objektivität und Reproduzierbarkeit (inter- bzw. intraindividuelle Reliabilität).

Da bisherige Studien gezeigt haben, dass inter- und intraindividuelle Reliabilität bei der Bewertung der Ergebnisqualität restaurativer Arbeiten nur unbefriedigend ist, wenn nach dem bisher allgemein üblichen „glance and grade“-System vorgegangen wird, ist es das Ziel der vorliegenden Studie festzustellen, inwieweit der Einsatz eines strukturierten Bewertungsbogens mit definierten Bewertungskriterien die Reliabilität der Benotung steigert.

Zu diesem Zweck wurden 30 im Phantomkurs der Zahnersatzkunde angefertigte Brücken von den Studierenden selber, einem Kommilitonen, zwei zahnmedizinischen Studenten des klinischen Studienabschnittes sowie jeweils zwei Zahnärzten aus der vorklinischen und der klinischen Kursbetreuung unabhängig voneinander und wiederholt beurteilt. Die Benotung der Arbeiten erfolgte dabei sowohl mittels Checkliste als auch unter Zuhilfenahme eines strukturierten Bewertungsbogens mit vorgegebenen Beurteilungskriterien.

Durch den Einsatz des detaillierten Bewertungsbogens konnte sowohl die Objektivität (interindividuelle Reliabilität) als auch die Reproduzierbarkeit (intraindividuelle Reliabilität) der Benotung nachweisbar gesteigert werden. Dabei wies die Gruppe der klinischen Kursassistenten sowohl beim inter- als auch beim intrapersonellen Vergleich die höchste Urteilskonkordanz auf. Die geringste interpersonelle Urteilkonkordanz wies die Gruppe der Studierenden selbst auf. Wich die Selbsteinschätzung der Studierenden ohne Bewertungsbogen bis zu eine Note von der Fremdbeurteilung ab, so konnte die Konkordanz mittels des Bewertungsbogens statistisch signifikant gesteigert werden.

Die vorliegenden Ergebnisse zeigen somit, dass durch den Einsatz eines strukturierten Bewertungsbogens die Zuverlässigkeit (Objektivität und Reproduzierbarkeit) der Benotung zahnmedizinischer Phantomarbeiten signifikant gesteigert werden kann und damit über den in vergleichbaren internationalen Studien angegebenen Korrelationswerten für andere Bewertungsverfahren liegt.

7 Literaturverzeichnis

1. *Baba K, Tsukiyama Y, Clark GT* (2000)
   Reliability, validity and utility of various occlusal measurement methods and techniques. J Prosthet Dent 83, 83.

2. *Barnes EJ, Pressey SL* (1929)
   The reliability and validity of oral examinations. School and Society 30, 719.

3. *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.

4. *Birkel P* (1978)
   Mündliche Prüfungen. Zur Objektivität und Validität der Leistungsbeurteilung. Kamp, Bochum.

5. *Dhuru VB, Rypel TS, Johnston WM* (1978)
   Criterion oriented grading system for preclinical operative dentistry laboratory course. J Dent Educ 42(9), 528-31.

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 effect 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 Proshtodontics: 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(9), 537-42.

10. Goldstein A (1958)

An inquiry into the value of rank grades in the medical course. J Med Educ 33, 193.

11. *Guild RE* (1966)

Questionaire studies at three Schools of dentistry. J Dent Educ 30 (4), 344-353.

12. *Heffer P, Holloway PJ, Rose JS, Swallow JN* (1965)

An investigation into dental undergraduate examing techniques. Br Dent J 118, 334-338.

13. *Hinkelman 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(7), 34-46.

15. *Ingenkamp K* (1995)

Die Fragwürdigkeit der Zensurengebung. Texte und Untersuchungsberichte. Beltz, Weinheim, 9. Aufl.

16. *Ingenkamp K* (1995)

Lehrbuch der Pädagogischen Diagnostik. Studienausgabe. Beltz, Weinheim, 4. Aufl.

17. *Jenkins SM et al.* (1998)

Evaluating undergraduate preclinical operative skill; use of a glance and grade marking system. J Dent 26(8), 679-84.

18. *Killip DE* (1965)

Role of discovery in learning manual skills. J Dent Educ 29, 63-70.

19. *King M, Bedi R* (1984)

The class II amalgam cavity in primary teeth. Dent Update 11, 413-419.

20. Leisen J, Mentges H (2000)

Klassische Formen der Leistungsmessung und Leistungsbeurteilung in der Schule. Studienseminar Koblenz).

<*URL: http://www.uni-koblenz.de/~odsssfg/seminar/uploads/modul17.pdf*>

21. *Lilley JD et al.* (1968)

Reliability of preclinical tests in operative dentistry. Br Dent J 3; 125(5), 194-7.

22. *Lissmann U* (1997)

Probleme und Möglichkeiten der Schülerbeurteilung. (Materialien für Lehre, Aus- und Weiterbildung, Bd. 8). Verlag Empirische Pädagogik, Landau ,2. Aufl.

23. *Mackenzie RS* (1973)

Defining clinical competence in terms of quality, quantity and need for performance criteria. J Dent Educ 37(9), 37-44.

24. *McDonald GT, Larson HD* (1985)

A system for developing and evaluating the clinical judgement of dental students. J Prosthet Dent 53(2), 265-266.

25. *Meetz HK, Bebeau MJ, Thoma SJ* (1988)

The validity and reliability of a clinical performance rating scale. J Dent Educ 52(6), 290-7.

26. *Natkin E, Guild RE* (1967)

Evaluation of preclinical laboratory performance: a systematic study. J Dent Educ 31(2), 152-161.

27. *Nebbe B et al.* (1998)

Interobserver reliability in quantitative MRI assessment of temporomandibular joint disk status. Oral Surg Oral Pathol Oral Radiol Endod 86, 746.

28. *O'Connor P, Lorey RE* (1987)

Improving interrater agreement in evaluating in dentistry by the use of comparison stimuli. J Dent Educ 42(4), 174-179.

29. *Robertello FJ, Pink, FE* (1997)

The effect of a training program on the reliability of examiners evaluating amalgam restorations. Oper Dent 22(2), 57-65.

30. *Sachs L, Henddrich J (2003)*

Angewandte Statistik. Anwendung statistischer Methoden. Springer, Berlin, 11. Aufl.

31. Salvendy G, Hinton WM, Ferguson GW, Cunningham PR (1973)

Pilot study on criteria in cavity preparation – facts or artefacts?. J Dent Educ 37, 27.

32. *Schiffler C* (2007)

Zur inter - und intraindividuellen Reliabilität der Beurteilung vorklinischer Zahnersatzarbeiten mittels Checkliste. [Dissertation]. Münster: Westfälische Wilhelms Universität.

33. *Shrout PE, Fleiss JL (1979)*

Intraclass correlation: Uses in assessing rater reliability.Psychological Bulletin, 86, 420-428.

34. Türp JC, Gerds Th, Schneider U (2002)

Variabilität bei der Benotung studentischer Arbeiten im vorklinischen Phantomkurs. Dtsch Zahnärztl Z 57, 526-531.

35. *Vanek HG* (1969)
Objektive Evaluation of Student Technic Products. J Dent Educ 33, 140-144.

36. *Vann WF, Machen JB, Hounshell PB* (1983)
Effects of criteria and checklists on reliability in preclinical evaluation. J Dent Educ 47(10), 671-675.

37. *Verhoeven JW, Cune MS, de Putter C* (2000)
Reliability of some clinical parameters of evaluation in implant dentistry. J Oral Rehabil 27, 211.

38. *Wahlund K, List T, Dworkin SF* (1998)
Temporomandibular disorders in children and adolescents: reliability of a questionnaire, clinical examination and diagnosis. J Orofac Pain 12, 42.

39. *Waugh D, Moyse CA* (1969)
Medical education. II. Oral examinations: a videotape study of the reproducibility of grades in pathologie. Can Med Assoc J 100, 635.

40. *Wirtz M* (2004)
Bestimmung der Güte von Beurteilereinschätzungen mittels der Intraklassenkorrelation und Verbesserung von Beurteilereinschätzungen. Rehabilitation 43, 384-389.

41. *Wirtz M, Casper F* (2002)
Beurteilerübereinstimmung und Beurteilerreliabilität. Hogreve-Verlag, Göttingen.

42. *Wolf B, von Bethlenfalvy E, Hassfeld S, Staehle HJ, Eickholz P* (2001)
Reliability of assessing interproximal bone loss by digital radiography: intrabony defects. J Clin Periodontol 28, 869.

43. *Wrabs KT, Kielbassa AM, Schulte Mönting J, Hellwig E* (1998)
Die Reproduzierbarkeit und Aussagekraft des Bißflügelbefundes. Dtsch Zahnärztl Z 53, 501.

44. *Zielinski J* (1981)

Mündliche Prüfungen und praktische Prüfungen. In: Twellmann, W. (Hrsg.): Handbuch Schule und Unterricht Band 42 Schule und Unterricht unter dem Aspekt der Didaktik unterrichtlicher Prozesse. Pädagogischer Verlag Schwann, Düsseldorf; 653-704.

Lebenslauf