Aus dem Dieter Scheffner Fachzentrum für medizinische Hochschullehre
und evidenzbasierte Ausbildungsforschung der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

DISSERTATION

Validierung eines empirisch basierten Beobachtungsinstruments für Unterrichtsqualität in der medizinischen Lehre

zur Erlangung des akademischen Grades

Doctor medicinae (Dr. med.)

vorgelegt der Medizinischen Fakultät

Charité – Universitätsmedizin Berlin

von

Anja Prescher

aus Berlin

Datum der Promotion: 09.12.2016

INHALTSVERZEICHNIS

1. ABSTRAKT

Einleitung

Die zuverlässige Beurteilung von Unterrichtsqualität ist notwendige Voraussetzung für die Evaluation und Weiterentwicklung medizinischer Ausbildung und damit letztendlich für eine hochqualitative Patientenversorgung. Entsprechende Qualitätsindikatoren können zudem Feedback für Dozierende auf eine nachvollziehbare Grundlage stellen. Ein vollständig empirisch basiertes Instrument zur Beurteilung von Unterrichtsqualität in der Medizin ist bisher nicht beschrieben. Daher wurden mit Hilfe empirisch basierter Kriterien für guten Unterricht aus dem Fachgebiet der Allgemeinen Pädagogik folgende zehn Kriterien entwickelt: Klare Strukturierung, Hoher Anteil echter Lernzeit, Lernförderliches Klima, Inhaltliche Klarheit, Sinnstiftende Kommunikation, Methodenvielfalt, Individuelles Fördern, Effektives Üben, Transparente Leistungserwartungen und Vorbereitete Unterrichtsumgebung. Die vorliegende Arbeit untersucht erstens die Frage, ob die extrahierten Kriterien Unterrichtsqualität in der universitären medizinischen Ausbildung valide erfassen können. Zweitens untersucht sie, da akademische Lehrevaluationen großteilig auf studentischen Beurteilungen beruhen, wie groß der Einfluss der einzelnen zehn Kriterien auf studentische Bewertungen von Unterrichtsqualität ist und drittens ob sich der Einfluss der Kriterien in verschiedenen Unterrichtsformaten oder zu verschiedenen Zeitpunkten im Studienverlauf unterscheidet.

Methodik

Zur Untersuchung dieser Fragestellungen wurde ein spezifischer Fragenbogen konzipiert, der Berlin Teaching Quality Questionnaire 10 (BTQ-10), mit dem eine Bewertung jedes der zehn Kriterien auf Basis einer inhaltlichen Kurzbeschreibung und eine Gesamteinschätzung der Unterrichtsqualität stattfand. Damit wurden 28 Unterrichtsveranstaltungen durch die jeweils teilnehmenden Studierenden in zwei verschiedenen klinischen Semestern (7. vs. 9. Fachsemester) sowie in zwei verschiedenen Unterrichtsformaten („Unterricht am Krankenbett“ vs. „Seminar“) an der Charité - Universitätsmedizin Berlin bewertet. Der Einfluss der zehn einzelnen Kriterien auf die Globalbewertung von Unterrichtsqualität wurde im Rahmen eines formativen Messmodells mit Hilfe einer ordinalen Regressionsanalyse bestimmt. Mittels Heterogeneous-Choice-Modellen wurde untersucht, ob zwischen den beiden Unterrichtsformaten und Zeitpunkten im Studienverlauf signifikante Unterschiede in den Einflussstärken der zehn Kriterien vorlagen.

Ergebnisse

Sieben Kriterien stellten signifikant positive Einflussfaktoren für die Globalbewertung von Unterrichtsqualität dar. Am stärksten war der Einfluss für das Kriterium „Sinnstiftende Kommunikation“. Dahingegen stellten die Kriterien „Lernförderliches Klima“, „Methodenvielfalt“ sowie „Vorbereitete Unterrichtsumgebung“ keine signifikanten Einflussfaktoren dar. Es zeigten sich keine Unterschiede zwischen den beiden Zeitpunkten im Studienverlauf. Zwischen den beiden untersuchten Unterrichtsformaten war ein signifikanter Unterschied in den Einflussstärken der Kriterien „Individuelles Fördern“ und „Inhaltliche Klarheit“ nachweisbar. Im Unterrichtsformat „Seminar“ wurde dem Kriterium „Individuelles Fördern“ ein signifikant höherer Stellenwert beigemessen als im Unterrichtsformat „Unterricht am Krankenbett“. Demgegenüber wies das Kriterium „Inhaltliche Klarheit“ im „Unterricht am Krankenbett“ einen signifikant stärkeren Einfluss auf.

Schlussfolgerung

Mit dem BTQ-10 steht erstmals ein auf die Anforderungen des deutschen Medizinstudiums zugeschnittenes valides Beobachtungsinstrument zur Beurteilung von Unterrichtsqualität aus studentischer Sichtweise zur Verfügung.

Introduction

Measuring the quality of teaching is a necessary prerequisite for the evaluation and development of medical education and thus for high-quality patient care. Corresponding quality indicators can make the feedback for teachers comprehensible. A completely empirically based instrument for the assessment of the quality of teaching in medicine has not yet been described. Ten empirically based criteria from the field of general pedagogy were developed: clear structure, amount of true learning time, climate facilitating learning, clarity of content, meaningful communication, diversity of methods, individual promotion, effective practice, transparent expectations and prepared setting. This study first assesses the question whether the extracted criteria validly capture the quality of teaching in medical education. Second, since academic teaching evaluation relies mostly on student assessment, it assesses the extent to which the ten criteria may have an impact on the students' ratings, and third, whether the impact of the criteria differs with various teaching formats or different semesters.

Methods

The Berlin Teaching Quality Questionnaire 10 (BTQ-10), in which each criterion can be rated based on a brief description of content and an overall rating of the quality of teaching, was developed to assess these questions. Participating students rated 28 courses in two different clinical semesters (7^{th} vs. 9^{th} semester) and two teaching formats ("bedside teaching" vs. "seminar") at the Charité - Universitätsmedizin Berlin. The impact of the ten criteria on the global rating of the quality of teaching was measured by a formative measurement model with an ordinal regression. By employing heterogeneous choice models, we assessed whether both teaching formats and semesters revealed significant differences in the impact of the ten criteria.

Results

Seven criteria were revealed to be significant positive influence factors for the global assessment of the quality of teaching. The strongest impact was found for the criterion "meaningful communication." By contrast, "climate facilitating learning," "diversity of methods," and "prepared setting" were not found to be significant influence factors. There was no difference between both semesters. A significant difference in effect was found between the assessed teaching formats concerning the criteria "individual promotion" and "clarity of content." "Individual promotion" was considered significantly more important in "seminar" than in "bedside teaching." In contrast, "clarity of content" was considered significantly more important in "bedside teaching."

Conclusion

The BTQ-10 provides a valid observational instrument to assess the quality of teaching from the student's perspective, tailored to the requirements of German medical education.

2. EINLEITUNG

2.1. Instrumente zur Beurteilung von Unterrichtsqualität in der medizinischen Ausbildung

Zwischen Unterrichtsqualität und dem Lernerfolg von Studierenden besteht ein positiver Zusammenhang [1-6]. Gute Lehre in der Medizin führt letztendlich auch zu einer Verbesserung der Patientenversorgung [7-10]. Daher ist die zuverlässige Beurteilung von Unterrichtsqualität notwendige Voraussetzung für die Evaluation und Weiterentwicklung medizinischer Ausbildung [11-15]. Mit Hilfe von Qualitätsindikatoren für guten Unterricht ist es zudem möglich, Feedback für Dozierende auf eine nachvollziehbare Grundlage zu stellen. Dieses spielt beispielsweise im Rahmen von Programmen zur Professionalisierung der Lehre in der Medizin eine wichtige Rolle und dient als wesentlicher Bestandteil in der Weiterbildung von Lehrern [16-21].

Zur Evaluation medizinischen Unterrichts wurden über die letzten Jahrzehnte zahlreiche Instrumente entwickelt und untersucht. Für den Zeitraum von 1966 bis 2010 führten Beckman et al. [22] und Fluit et al. [8] in zwei Reviews insgesamt 56 Artikel mit 34 Messinstrumente auf. Vaughan [23] hat die Übersicht bis 2013 fortgesetzt und zusammenfassend 67 Artikel gefunden, die sich mit der Erfassung von Qualität in der medizinischen Ausbildung beschäftigen. In einer aktuellen Übersichtsarbeit zur Evaluation im Medizinstudium geben Schiekirka et al. [24] einen Überblick über bestehende Erhebungsinstrumente. Tabelle 1 listet häufig verwendete Instrumente auf.

Die Instrumente unterscheiden sich in mehreren Punkten. Dazu gehören der untersuchte Abschnitt im Verlauf der medizinischen Ausbildung, das heißt die Zeit der medizinischen Ausbildung während des Studiums (undergraduate medical education) [25-32] oder der Bereich der medizinischen Weiterbildung im Rahmen der Facharztausbildung (graduate medical education) [33-37]; die Wahl der klinischen Umgebung, d.h. der stationärer Bereich [25, 26, 31, 32, 34, 38], die Ambulanz [39, 40] oder die Notaufnahme [37]; und der Personenkreis, der die Evaluation vornimmt, das heißt Studierende [40-44], Assistenzärzte [10, 33, 35, 37, 45], geschulte nicht-medizinische Beobachter [46, 47] oder Peers [34]. Wie auch in Tabelle 1 ersichtlich, stammt der Großteil der bisher publizierten Messinstrumente aus dem angelsächsischen Sprachraum. Die Übertragbarkeit auf den Kontext des deutschen Medizinstudiums ist dabei eingeschränkt [24].

Tabelle 1: Überblick über häufig benutze Messinstrumente zur Erfassung von Unterrichtsqualität

Instrument , Land, Erscheinungsjahr, Ref.	I	S	Bezeichnung der Skalen
Clinical Teaching Assessment Form (CTAF), USA, 1981, [27]	9	8	knowledge and analytical ability; clarity and organization; enthusiasm and stimulation; ability to establish rapport; ability to involve student in learning experiences; ability to give direction and feedback; ability to demonstrate clinical skills and procedures; accessibility
Clinical Tutor Evaluation Questionnaire (CTEQ), Kanada, 1993, [48]	25	7	attitude to teaching; humanistic orientation; perceived subject matter expertise; teaching skills; problem-solving emphasis; student-centered teaching strategy; active student participation
Teaching Effectiveness Score (TES), USA, 1996, [38]	28	4	attitude towards teaching medical students; ability to provide useful feedback; ability to challenge thinking; ability to communicate and answer questions clearly
Stanford Faculty Development Program (SFDP26), USA, 1998, [31]	25	7	learning climate; control of session; communication of goals; understanding and retention; evaluation; feedback; self-directed learning
Medical Instructional Quality (MedIQ), USA, 1999, [39]	25	4	role of the preceptor in facilitating learning; role and context of the clinical environment; opportunities available to learn; active involvement by the learner in the care of patients
Clinical Teaching Effectiveness Instrument (CTEI), USA, 2000, [49]	15	-	
Mayo Teaching Evaluation Form (MTEF), USA, 2003, [34	28	7	establishing a positive learning climate; control of teaching session; communication of goals; enhancing understanding and retention; evaluation; feedback; self-directed learning
Student Evaluation of Teaching in Outpatient Clinics (SETOC), USA/Pakistan, 2007, [40]	15	5	establishing learning-milieu; clinical teaching; general-teaching; clinical-competence; global rating
Maastricht Clinical Teaching Questionnaire (MCTQ), Niederlande, 2010, [43]	16	5	modeling; coaching; exploration; articulation; safe learning environment
SFDP26-German, Deutschland, 2011, [50]	25	7	Etablieren des Lernklimas; Leitung einer Lehreinheit; Zielkommunikation; Fördern von Verstehen und Behalten; Evaluation; Feedback; Fördern selbstbestimmten Lernens

Ref.: Referenz, I: Anzahl der benutzten Items, S: Anzahl der Skalen, die von den Items reflektiert werden

Eines der am häufigsten benutzen Instrumente ist der SFDP26. Er stammt aus dem Stanford Faculty Development Program (SFDP). Dabei handelt es sich um ein Fakultätsentwicklungsprogramm, das Anfang der achtziger Jahre an der Stanford University, Palo Alto, USA, etabliert wurde. Dabei wurden Dozierende im Unterricht beobachtet und deren Stärken und Schwächen in der Lehre in sieben Kernkompetenzen differenziert. Dabei handelt es sich um „learning climate“, „control of session“, „communication of goals”, “understanding and retention”, “evaluation”, “feedback” und “self-directed learning” [31, 32, 51, 52]. Der SFDP26 setzt sich zusammen aus 25 Items, die die sieben Kernkompetenzen operationalisieren, sowie einem zusätzlichen Item zur „overall teaching effectiveness“. Es existiert eine validierte deutsche Übersetzung, der SFDP26-German [50].

Einige Instrumente basieren auf einem einzigen Item zur Gesamtbeurteilung des Unterrichts [45, 51, 53,45, 53]. Diese bieten zwar im Rahmen von fakultätsweiten Evaluationsprogrammen und der Evaluation mehrerer Lehrveranstaltungen und Lehrer sowie Zeitdruck eine mögliche Lösung, beinhalten jedoch kaum Anknüpfungspunkte für die Verbesserung von Unterricht [27, 45, 51, 53-56].

Viele Instrumente beschäftigen sich weniger mit der Entwicklung von beobachtbaren Qualitätsindikatoren für guten Unterricht als vielmehr mit schlecht messbaren, deskriptiven Eigenschaften guter klinischer Lehrer [15, 54, 57-65]. Sutkin et al. haben in ihrem Review aus 68 Artikeln von 1966 bis 2008 480 Eigenschaften eines guten klinischen Lehrers gesammelt und diese in 49 Themenkomplexe in die drei Hauptkategorien „physician characteristics“, „teacher characteristics“ und „human characteristics“ gruppiert [64].

2.2. Beurteilung von Unterrichtsqualität durch Studierende

Die Evaluation von Lehrveranstaltungen durch Studierende ist heutzutage an fast allen Hochschulen im In- und Ausland Standard [6, 66, 67]. Obwohl die Gültigkeit studentischer Beurteilungen von Lehrveranstaltungen angezweifelt wurde, zeigt der aktuelle Forschungsstand, dass über Veranstaltungen gemittelte studentische Beurteilungen multidimensional, messgenau und stabil sind. Sie stellen primär eine Funktion des Lehrverhaltens des Dozierenden und nicht der Kursthemen dar. Sie sind valide hinsichtlich vieler Indikatoren effektiver Lehre, zum Beispiel Lerntests als Leistungsmaß sowie Fremdbeurteilungen, relativ unbeeinflusst von einer Vielzahl von potentiellen Verzerrungsvariablen wie Geschlecht und Intelligenz und besser geeignet die Qualität von Lehrveranstaltungen zu evaluieren als die Beurteilungen der Dozierenden selbst [6, 66,68, 69].

Boerboom et al. untersuchten die Abhängigkeit der Ergebnisse des Maastricht Clinical Teaching Questionnaire (MCTQ) von Eigenschaften der Studierenden und Dozierenden, die nichts mit der eigentlichen Lehrleistung zu tun haben. Dazu zählten die Erfahrung des Dozierenden in der klinischer Lehre, das Geschlecht des Dozierenden und der Studierenden, die Abteilung des Dozierenden, das Studienjahr der Studierenden und die Lehrqualifikation des Dozierenden. Alle zeigten keinen signifikanten Einfluss auf die Ergebnisse des MCTQ. Dieses bestätigt die Validität studentischer Bewertungen [70]. Zuvor wurde dieses sonst nur für studentische Evaluationen außerhalb der medizinischen Ausbildung gezeigt [66]. Albanese et al. legten diesbezüglich dar, dass die Bewertungen von geschulten nicht-medizinischen Beobachtern die Bewertungen von Studierenden im Rahmen von medizinischen Lehrveranstaltungen voraussagen konnten [71].

Des Weiteren wurde die Frage aufgeworfen, ob Studierende nicht nur einzelne Aspekte von Unterrichtsqualität bewerten können, sondern auch in der Lage sind, eine Globalbeurteilung vorzunehmen. Marriott et al. [51] sowie Williams et al. [45, 53] haben diesbezüglich gezeigt, dass die globale Einschätzung von Unterrichtsqualität im Rahmen eines Items reliabel, konsistent und valide ist. Sie widersprechen damit dem Argument, dass sogenannte „global ratings“ dem Halo-Effekt unterliegen und vielmehr die Beliebtheit oder Persönlichkeit des Lehrers erfassen [27, 72]. Marriott und Litzelman [51] weisen in diesem Zusammenhang darauf hin, dass, wenn Studierende am Ende eines Fragebogens aufgefordert werden ein „global rating“ abzugeben, sie sich vorher schon mit den einzelnen Qualitäten des Unterrichts beschäftigt haben und somit die Gesamteinschätzung das vollständige Instrument erfassen kann.

2.3. Validität von Messinstrumenten

Für eine zuverlässige Erfassung von Unterrichtsqualität, eine inhaltlich fundierte Interpretation der Ergebnisse und die Ableitung von Konsequenzen sind valide Messinstrumente unabdingbar [14, 24, 27, 38, 39, 42, 43, 55, 73]. Validität beschreibt die „Gültigkeit“ eines Messinstruments. Sie liegt vor, wenn das Instrument das Konstrukt, das es zu messen vorgibt, auch tatsächlich erhebt. Traditionell werden Inhaltsvalidität, Kriteriumsvalidität und Konstruktvalidität unterschieden. Inhaltsvalidität ist gegeben, wenn das Konstrukt hinreichend präzise durch den Inhalt der Items definiert wird und in seinen wichtigsten Aspekten vollständig erfasst ist. Kriteriumsvalidität liegt vor, wenn die Messergebnisse ein mit dem gemessenen Konstrukt zusammenhängendes externes Kriterium vorhersagen können. Konstruktvalidität gibt an, inwieweit ein Instrument das zu erfassende Konstrukt so misst, dass es mit bestehenden Konstruktdefinitionen und Theorien übereinstimmt. Sie umfasst die konvergente und

diskriminante Validität. Konvergente Konstruktvalidität besagt, dass ein Merkmal valide gemessen wird, wenn die Ergebnisse mit jenen eines Messinstruments übereinstimmen, welches das gleiche Konstrukt zu messen beansprucht. Diskriminante Validität liegt vor, wenn die Ergebnisse zweier Messinstrumente, die unterschiedliche jedoch ähnliche Konstrukte messen, wenig übereinstimmen [74-78]. In neuerer Zeit wird von einem integralen Validitätskonzept ausgegangen, das die oben genannten Konzepte miteinander verbindet. Validität wird in diesem Sinne durch verschiedene Quellen der Evidenz für die Zulässigkeit von Schlüssen belegt. Diese umfassen Evidenz auf der Basis der Inhalte, der Antwortprozesse, der internen Struktur, der Relation zu anderen Merkmalen und der Konsequenzen [14, 79].

Insofern ist es wünschenswert, wenn die Herkunft von Qualitätsindikatoren bekannt ist und dass für diese eine empirische Basis besteht. Der Ursprung der einzelnen Items und das empirische Fundament der bisher publizierten Instrumente zur Beurteilung von Unterrichtsqualität ist jedoch lückenhaft [8, 22, 25, 39, 42, 43, 55, 73]. Zahlreiche Instrumente bauen aufeinander auf oder verwenden verschiedene Items bereits psychometrisch untersuchter Instrumente [27, 29, 31, 34, 35, 37, 40, 42, 44, 46, 49, 80, 81]. So basiert zum Beispiel der Mayo Teaching Evaluation Form (MTEF) auf dem SFDP26 [34].

Ein ausschließlich empirisch basierter Bezug der Indikatoren zur Erfassung von Unterrichtsqualität für die medizinische Ausbildung ist bisher in der Literatur nicht beschrieben. Daher haben Breckwoldt et al. in Anlehnung an empirisch basierte und gut beobachtbare Qualitätsmerkmale aus der allgemeinen Pädagogik mit Hilfe einer umfassenden Literaturanalyse zehn empirisch basierte Kriterien zur Erfassung von Unterrichtsqualität in der medizinischen Ausbildung extrahiert [82]. Die Basis für diese Merkmale hat Slavin 1994 mit einem Konstrukt aus vier Kategorien für guten Unterricht, dem QAIT-Modell (Quality, Appropriateness, Incentive, Time), gelegt [83]. Als weiteren Grundstein erstellte Brophy 1999 für das International Bureau of Education der UNESCO ein Konstrukt für Unterrichtsqualität aus 12 Merkmalen, die er auf zahlreiche empirische Studien stützte [84]. Diese 12 Kriterien bildeten die Basis für den Kriterienkatalog für guten Unterricht von Helmke [85] und Meyer [86]. Breckwoldt et al. extrahierten in Bezug auf diese vier Autoren folgende zehn Kriterien: „Klare Strukturierung“, „Hoher Anteil echter Lernzeit“, „Lernförderliches Klima“, „Inhaltliche Klarheit“, „Sinnstiftende Kommunikation“, „Methodenvielfalt“, „Individuelles Fördern“, „Effektives Üben“, „Transparente Leistungserwartungen“ und „Vorbereitete Unterrichtsumgebung“.

2.4. Zehn empirisch basierte Kriterien zur Erfassung von Unterrichtsqualität

Die extrahierten empirisch basierten Kriterien lassen sich wie folgt charakterisieren:

- Klare Strukturierung

Der Kernbereich des Kriteriums „Klare Strukturierung“ stellt die Kompatibilität von Zielen, Inhalten und verwendeten Methoden des Unterrichts dar. Der methodische Gang soll nachvollziehbar und kohärent sein, um den roten Faden des Unterrichts allzeit zu erkennen. Der Unterricht soll einem didaktischen Grundrhythmus folgen, der sich aus Vorbereitung, Erarbeitung sowie Übung und Ergebnissicherung zusammensetzt. Des Weiteren sollen die Rollen und Aufgaben während des Unterricht klar und präzise definiert sein.

- Hoher Anteil echter Lernzeit

Das Kriterium „Hoher Anteil echter Lernzeit“ umfasst die tatsächliche Lernzeit sowie das Lerntempo während des Unterrichts. Unterschiedliche Faktoren führen zu einer Reduktion der zur Verfügung stehenden Unterrichtszeit. Dazu zählen organisatorische Aktivitäten wie das Kontrollieren der Anwesenheit und die Überprüfung der Lehrmaterialen, Störungen des Unterrichts, Unklarheiten zum Beispiel zu Aufgaben, Verspätungen des Dozierenden sowie ein deutlich früheres Beenden des Unterrichts ohne Erreichen der Unterrichtsziele. Das interindividuell optimale Lerntempo während des Unterrichts kann durch implizite und explizite Rückmeldungen der Studierenden festgestellt und vom Dozierenden entsprechend angepasst werden. Zu langsames Lerntempo führt zu Langeweile, Abnahme der Motivation und Aufmerksamkeit sowie Störungen des Unterrichts. Ein zu schneller Fortgang im Unterrichtsgeschehen hingegen führt zu Überforderung, Unklarheiten und in der Folge ebenso einem Verlust von Motivation und Aufmerksamkeit. Beides ist mit einer Reduktion der echten Lernzeit verbunden.

- Lernförderliches Klima

Ein unterstützendes Unterrichtsklima trägt stark zum Lernerfolg bei. Es soll frei von Diskriminierung und geprägt von Wertschätzung sein sowie Dozierende und Studierende in die Lage versetzen im Sinne einer Lerngemeinschaft die Verantwortung zur Erreichung der Lernziele zu teilen. Durch eine fördernde, aufrichtige und gerechte Haltung des Dozierenden ist die Etablierung einer Fehlerkultur möglich und Fehler werden als integraler Bestandteil des Lernprozesses angesehen. Durch diese Grundhaltungen von Seiten des Dozierenden ist eine Identifikation der Studierenden mit dem Dozierenden im Sinne eines Vorbildes möglich. Ein

lernförderliches Klima ist unter anderem an der effektiven Verwendung von Regeln und wenigen Störungen und Ablenkungen festzumachen, stärkt die Selbstwirksamkeit und Interessenbildung der Studierenden und führt zu einer höheren Bereitschaft zur aktiven Teilnahme am Unterricht.

- Inhaltliche Klarheit

Neue Lerninhalte sollen klar aufeinander aufgebaut und in Beziehung zueinander gesetzt werden. Damit für den Dozierenden Vorwissen und Kompetenzen der Studierenden als Anknüpfungspunkte ersichtlich sind, ist zunächst eine Lernstandsanalyse notwendig. Der Dozierende soll zu entsprechenden Zeitpunkten während des Unterrichts den Lernfortschritt der Studierenden durch das Einholen von Rückmeldungen monitoren. Damit vergewissert er sich, dass ein thematischer Schritt gedanklich von den Studierenden nachvollzogen wurde oder praktisch umgesetzt werden kann, bevor er zum nächsten Schritt übergeht. Neben eindeutigen und verständlichen Fragen und Aufgaben spielen die fachliche Richtigkeit und korrekte fachsprachliche Ausdrucksweise des Dozierenden sowie der Unterrichtsinhalte eine wichtige Rolle für die inhaltliche Klarheit. Für eine hohe inhaltliche Klarheit sorgt des Weiteren eine klare, prägnante und verbindliche Ergebnissicherung. Die Ergebnissicherung stellt einen zeitkritischen Teil des Unterrichts dar. Eine zu kurz gewählte Ergebnissicherung kann wichtige Ergebnisse vernachlässigen, eine zu lange Ergebnissicherung kann infolge von Redundanz zu einem Aufmerksamkeitsdefizit der Studierenden führen.

- Sinnstiftende Kommunikation

Da Lernende Unterrichtsinhalten und Lernprozessen immer eine persönliche, affektiv aufgeladene und nachfolgend als Assoziationshilfe dienende Bedeutung geben, ist es sehr wichtig, Lernprozessen und Unterrichtsinhalten möglichst positive Bedeutungen zu geben. Der Dozierende kann diese Sinngebung durch die Studierenden beeinflussen, indem er unter anderem den individuellen Nutzen der Lerninhalte für den einzelnen Studierenden herausstellt. Die Reaktion des Dozierenden auf Unterrichtsbeiträge oder Leistungen in Übungsphasen einzelner Studierender leistet ebenso einen relevanten Beitrag zur sinnstiftenden Kommunikation. Feedback hat einen wichtigen kommunikativen Aspekt. Um vom Lernenden angenommen zu werden und Änderungen bewirken zu können, sollte Feedback fair und passend kommuniziert werden. Die Etablierung und Weiterentwicklung einer Feedbackkultur ermöglicht Studierenden auch zukünftige Lernumgebungen besser zu gestalten. Des Weiteren trägt sinnstiftende Kommunikation zur Interessenbildung sowie Motivation bei Lernenden und damit zu einer besseren Leistung auf dem entsprechenden Gebiet bei.

- Methodenvielfalt

Um der Heterogenität von individuellen Voraussetzungen, Lerntypen sowie Interessen der Studierenden gerecht zu werden und verschiedene Lerninhalte vermitteln zu können, ist die Anwendung einer Vielfalt von Unterrichtsmethoden notwendig. In der Medizin existieren zahlreiche Unterrichtsmethoden, die häufig und erfolgreich angewandt werden wie zum Beispiel bed-side teaching und Problem-orientiertes-Lernen. Das Kriterium Methodenvielfalt bewertet weder die Anzahl wechselnder Methoden noch die Überlegenheit einzelner Methoden, sondern beschreibt die sinnvolle Auswahl und Abstimmung an Methoden für den entsprechenden Unterrichtsgegenstand. Methoden umfassen sowohl Medien und Unterrichtsmaterialien als auch Sozialformen und Unterrichtsmodelle. Des Weiteren umfasst das Kriterium „Methodenvielfalt" auch die Vermittlung von Lernstrategien.

- Individuelles Fördern

Das Kriterium „Individuelles Fördern" betrachtet, ob alle Studierende sich trotz verschiedener individueller Voraussetzungen und Möglichkeiten innerhalb einer Unterrichtseinheit Wissen und Fertigkeiten aneignen können. Dieses erfordert ein kontinuierliches Monitoring des Lernfortschritts durch den Dozierenden, um Defizite sofort zu erkennen und lernfördernde Schritte einzuleiten. Dazu sollten unter anderem möglichst alle Studierenden in Fragen, Diskussionen und Übungsphasen einbezogen werden. Die Lernenden sollten jede mögliche Unterstützung durch den Dozierenden erhalten, um vom Unterricht zu profitieren. Welcher Lernende dabei welche Art von Unterstützung durch den Lehrer benötigt, sollte dieser erkennen und zeitnah umsetzen. Individuelle Förderung soll Hilfe zur Selbsthilfe sein.

- Effektives Üben

Übungsphasen im Unterricht haben drei grundsätzliche Ziele: Transfer von Wissen oder Können, Automatisierung und Qualitätssteigerung. Äußere Bedingungen für eine hohe Qualität von Übungsphasen sind einerseits die Motivation der Studierenden zu Wiederholung und Anwendung des Gelernten und andererseits ausreichend Zeit zum Üben. Damit Übungsphasen qualitativ hochwertig sind, sollen sie zum Lernstand der Studierenden und zum Gegenstand des Unterrichts passen und den Studierenden Erfolge ermöglichen. Während des Unterrichts sollen Übungen für die Studierenden nachvollziehbar in den Kontext eingebettet sein. Effektives Üben erfordert zeitnahes Feedback. Dieses soll die Lernenden informieren und konkrete Beobachtungen und ggf. konkrete Änderungsvorschläge enthalten statt zu werten. Feedback

kann die Studierenden dabei unterstützen, Lernziele zu erreichen, eigene Lernfortschritte wahrzunehmen, Fehler zu korrigieren und zukünftig zu vermeiden.

- Transparente Leistungserwartungen

Transparente Leistungserwartungen sollen die Studierenden motivieren und unterstützen, ihre Lernziele zu erreichen. Damit Leistungserwartungen transparent werden, sollen sie klar, realistisch und für die Studierenden nachvollziehbar sein. Dem Dozierenden sollen sie als solche bewusst sein und er soll sie ggf. an den Lernstand und den Unterrichtskontext anpassen und verständlich kommunizieren. Durch wiederholte Rückmeldungen des Dozierenden über den individuellen Lernfortschritt der Studierenden bleiben die Leistungserwartungen für Dozierende und Studierende stets präsent.

- Vorbereitete Unterrichtsumgebung

Das Kriterium „Vorbereitete Unterrichtsumgebung“ umfasst die räumlichen und technischen Ressourcen für guten Unterricht. Verwendete Materialen und Medien sollen vorhanden sowie funktionsfähig sein und der Dozierenden im Umgang mit Ihnen vertraut. Es soll ein geeigneter Unterrichtsraum zur Verfügung stehen. Beim Einsatz von Patienten während des Unterrichts ist ebenfalls eine Vorbereitung erforderlich, geeignete Patienten müssen zunächst identifiziert und ihr Einverständnis zur Teilnahme am Unterricht eingeholt werden.

2.5. Operationalisierung eines theoretischen Konstrukts

Ein theoretisches Konstrukt stellt a priori eine nicht direkt messbare Größe dar. Man bezeichnet sie auch als latente Variabel. Um eine latente Variabel indirekt zu erfassen, können Konstrukte mit Hilfe reflektiver und formativer Messmodelle operationalisiert werden. Die Unterscheidung betrifft die Kausalität des Zusammenhangs zwischen der latenten Variabel und den dazugehörigen manifesten Variablen. Die manifesten Variablen werden als Indikatoren bzw. in Fragebögen als Items bezeichnet. In reflektiven Messmodellen verursacht die latente Variabel die Indikatoren. In formativen Messmodellen sind im Gegensatz dazu die Indikatoren die Verursacher der latenten Variabel [87-95].

2.5.1. Reflektive Messmodelle

In reflektiven Messmodellen werden Konstrukte als Ursache beobachtbarer Indikatoren interpretiert [90]. Gemäß der klassischen Testtheorie lassen sich unendlich viele Indikatoren für ein reflektives Konstrukt heranziehen. Zur Erfassung des Konstrukts wird eine Stichprobe an Indikatoren aus dem theoretischen Itemuniversum gezogen [89, 92, 96, 97]. Zur Optimierung der

Reliabilität kann über einen Skalenbereinigungsprozess eine Selektion erfolgen und messfehlerbehaftete Indikatoren ausgeschlossen werden ohne den konzeptionellen Rahmen des Konstrukts zu verändern [77, 97]. Da die Ausprägungen der Indikatoren kausal durch die latente Variabel verursacht werden, geht eine Veränderung der Konstruktausprägung mit einer Veränderung aller Indikatorausprägungen einher [91, 92, 95]. Da jeder einzelne Indikator das gesamte Konstrukt wiederspiegelt, sind hohe Korrelationen zwischen den Indikatoren zu erwarten [95]. Wären die Indikatoren perfekte Messungen des Konstrukts, so wiesen sie untereinander einen Korrelationskoeffizienten von 1 auf [90, 91].

Die Spezifikation eines reflektiven Messmodells erfolgt gemäß der Formel [87]:

$$x_i = \lambda_i \xi + \delta_i \ (i = 1, \dots, n)$$

mit x_i : der i-te reflektive Indikator,

ξ : die latente Variabel,

λ_i : die Faktorladung,

δ_i : der Messfehler auf Indikatorebene.

In diesem System linearer Gleichungen ist jeder reflektive Indikator x_i als ein mit einer Ladung λ_i gewichtetes Abbild der latenten Variabel ξ darstellbar [91, 95]. Jeder reflektive Indikator ist mit einem systematischen und zufälligen Messfehler behaftet. Abbildung 1 stellt ein reflektives Messmodell dar. Als Methoden zur Beurteilung der Gütekriterien Reliabilität und Validität des reflektiven Messmodells können unter anderem Cronbachs alpha als Maß für die interne Konsistenz und die konfirmatorische Faktorenanalyse herangezogen werden [77, 87, 91, 92].

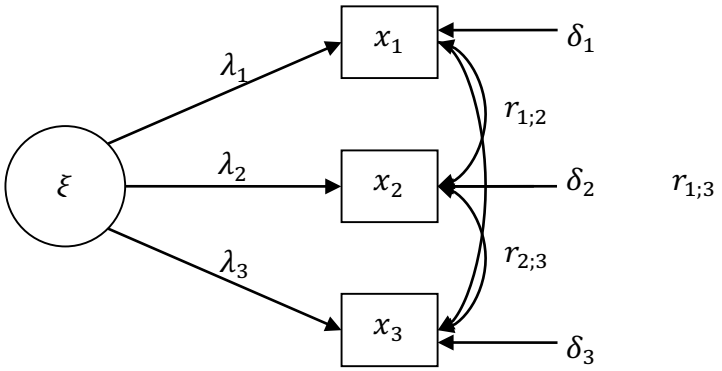

Abbildung 1: Reflektives Messmodell [87]
$\boldsymbol{\xi}$: latente Variabel, $\boldsymbol{\lambda}$: Faktorladung, $\boldsymbol{x}$: reflektiver Indikator, $\boldsymbol{\delta}$: Messfehler auf Indikatorebene, $\boldsymbol{r}$: Korrelation zwischen den Indikatoren

2.5.2. Formative Messmodelle

In formativen Messmodellen sind die Indikatoren Verursacher der latenten Variabel [87, 89-91, 95, 98, 99]. Veränderungen eines einzelnen Indikators führen zu einer Veränderung der latenten Variabel. Eine Veränderung der latenten Variabel ist nicht notwendigerweise mit einer Veränderung aller oder auch nur einiger Indikatoren verbunden [91, 93]. Darüber hinaus nehmen die Indikatoren nicht zwangsläufig in gleich starkem Maße Einfluss auf die latente Variabel. Jede manifeste Variabel erhält ein individuelles Gewicht. Je höher das Gewicht eines Indikators ausfällt, desto stärker ist dessen Beitrag für die inhaltliche Bestimmung des ihm zugeordneten formativen Konstrukts. Somit ist es möglich Einflussstärken der einzelnen Indikatoren zu identifizieren und diese zur Veränderung der Konstruktausprägung in die geforderte Richtung anzupassen. Innerhalb eines formativen Messmodells kann das latente Konstrukt als eine gewichtete Zusammensetzung seiner Indikatoren betrachtet werden [89-91, 98].

Die latente Variabel η ist als Linearkombination der Indikatoren dargestellt, was im Wesentlichen dem klassischen multivariaten Regressionsmodell entspricht [91, 95]. Der Messfehler wird der latenten Variabel selbst zugeordnet. Der sich aus den Indikatorwerten ergebende Konstruktwert stimmt nicht mit dem wahren Konstruktwert überein, da die Indikatoren das Konstrukt nie vollständig abbilden können [90, 95].

Die Spezifikation eines formativen Messmodells erfolgt gemäß der Formel [87, 89, 94, 95]:

$$\eta = \gamma_1 x_1 + \gamma_2 x_2 + \ldots + \gamma_n x_n + \zeta = \sum_{i=1}^{n} \gamma_i x_i + \zeta$$

mit η : die latente Variabel,

γ_i: das i-te Gewicht,

x_i der i-te formative Indikator,

ζ : der Messfehler auf Ebene der latenten Variabel.

Abbildung 2 stellt ein formatives Messmodell dar.

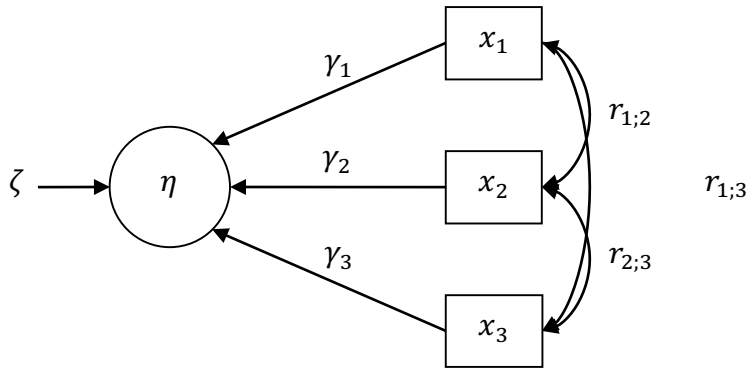

Abbildung 2: Formatives Messmodell [87]
η: latente Variabel, γ: Gewicht, x: formativer Indikator, ζ: Messfehler auf Ebene der latenten Variabel, r: Korrelation zwischen den Indikatoren

Je nach Konstrukt kann die Bewegung eines Indikators durch die Gegenbewegung eines anderen Indikators substituiert werden, so dass die Konstruktausprägung unverändert bleibt [90]. Ob und inwieweit sich bei Veränderung eines Indikators gleichzeitig auch die anderen Indikatoren verändern, ist durch die Korrelationen zwischen den Indikatoren bestimmt [91]. Die Indikatoren müssen nicht untereinander korrelieren, eine hohe Korrelation ist jedoch durchaus möglich [89, 90, 95, 100]. Somit verbietet sich auch die Eliminierung eines einzelnen Indikators auf Grundlage von Kriterien der klassischen Testtheorie wie der internen Konsistenz sowie die Anwendung der Faktorenanalyse [89, 94, 95]. Die Beurteilung der Modellgüte ist somit nicht wie bei reflektiven Messmodellen möglich [89, 91-95, 101, 102].

2.5.2.1. Operationalisierung und Gütebeurteilung eines formativen Messmodells

Die Operationalisierung eines formativen Messmodells findet in mehreren Schritten statt und ist in Abbildung 3 dargestellt.

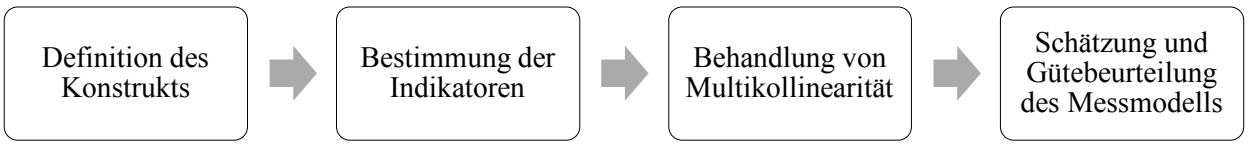

Abbildung 3: Operationalisierung eines formativen Messmodells [88, 89, 92, 95]

Zunächst erfolgt eine präzise Konstruktdefinition, die alle relevanten Facetten des Konstrukts umfasst [89, 92, 103]. Dazu ist es möglich Fallstudien, Interviews sowie Befragungen von Experten durchzuführen [89-92]. Darauffolgend wird die Menge der formativen Indikatoren, die einen Einfluss auf die Ausprägung des Konstrukts ausüben, identifiziert. Eine Nichtberücksichtigung einzelner Indikatoren führt zu einer Änderung des konzeptionellen Inhalts

der latenten Variabel [89]. Die Indikatoren werden einer Qualitätsprüfung unterzogen. Diese umfasst die Verständlichkeit sowie die inhaltliche Validität eines Indikators, d.h. ob jeder Indikator die inhaltliche Facette, die er erfassen soll, auch wirklich trifft. Dies kann durch eine Befragung von Experten oder repräsentativen Ratern erfolgen [89-91, 102].

Da formative Messmodelle auf Basis der multiplen Regressionsanalyse berechnet werden können, stellt Multikollinearität zwischen formativen Indikatoren ein Problem dar [89, 90, 94, 95]. Als Multikollinearität wird der Grad der linearen Abhängigkeit zwischen zwei oder mehreren Indikatoren bezeichnet [101]. Beim Vorliegen von hoher Multikollinearität kommt es zu ungenauen Schätzungen der Gewichte der entsprechenden Indikatoren [89, 95, 103]. Zur Prüfung und Behandlung von Multikollinearität stehen verschiedene Methoden zur Verfügung. Eine Elimination multikollinearer Indikatoren ist umstritten, da sie mit dem Verlust einer relevanten inhaltlichen Facette des Konstrukts verbunden ist [87, 89, 90, 92, 94, 103].

Um eine Schätzung und Gütebeurteilung eines formativen Messmodells vornehmen zu können, muss die latente Variabel in ein größeres Strukturmodell eingebunden werden [89, 90, 94, 95, 104]. Die Überprüfung der Hauptgütekriterien Reliabilität und Validität kann nicht wie bei reflektiven Modellen gemäß der klassischen Testtheorie erfolgen. Die Bestimmung der Reliabilität, die bei reflektiven Indikatoren auf den Korrelationen untereinander und internen Konsistenzmaßen beruht, ist bei formativen Indikatoren nicht möglich. Wenn gewährleistet werden kann, dass der Kontext bei zwei zu verschiedenen Messzeitpunkten durchgeführten Messungen identisch ist, ist es möglich die Retest-Reliabilität zu bestimmen [88, 95, 105].

Das Vorgehen zur Beurteilung der Validität formativer Messmodelle wird kontrovers diskutiert. Im Vordergrund steht die Überprüfung der Inhaltsvalidität durch Expertenurteile [89-92, 94, 95, 102, 103, 105, 106]. Falls ein Konstrukt sowohl mit formativen wie auch reflektiven Indikatoren operationalisiert werden kann, kann in einem Zwei-Konstrukt-Modell die reflektive Skala als abhängige Variabel zur Beurteilung der Konstruktvalidität dienen. Ist neben der formativen Erfassung auch eine direkte Beobachtung des Konstrukts möglich, kann die direkte Messung als abhängige Variabel für das zu validierende formative Maß dienen. Bei Vorhandensein von mehr als einem Konstrukt können diese mit Strukturgleichungsmodellen geschätzt und validiert werden [107]. Liegen nur formative Indikatoren im Rahmen eines Konstrukts vor, empfiehlt sich eine multiple Regressionsanalyse. Dabei ist das Bestimmtheitsmaß von Bedeutung, um eine Aussage über die Konstruktvalidität zu treffen. Die Regressionskoeffizienten stellen die Gewichte der einzelnen formativen Indikatoren dar [103].

2.5.3. Entscheidungskriterien zur Wahl eines reflektiven oder formativen Messmodells

Fehler in der Wahl der Indikator-Konstrukt-Beziehung führen zu Verzerrungen der Schätzergebnisse und fehlerhaften Untersuchungsergebnissen. Inwiefern das abzubildende Konstrukt ein formatives oder reflektives Messmodell erfordert, muss primär aus der Theorie erfolgen [89, 91, 92, 99, 103]. Die in Tabelle 2 dargestellten Entscheidungskriterien können dabei herangezogen werden [88, 90-93].

Tabelle 2: Entscheidungskriterien zur Wahl eines reflektiven oder formativen Messmodells

	Reflektives Messmodell	Formatives Messmodell
Kausalität zwischen latenter Variable (LV) und Indikatoren	• Die Indikatoren ergeben sich als Folge (Konsequenz) aus der LV. • Die Indikatoren sind als Manifestation der LV anzusehen. • Die LV steht zeitlich gesehen vor den Indikatoren.	• Die Indikatoren stellen die Ursache der LV dar. • Die Indikatoren bestimmen in ihrer Kombination den Inhalt der LV. • Die Indikatoren stehen zeitlich gesehen vor der LV.
Elimination und Austauschbarkeit der Indikatoren	• Einzelne Indikatoren lassen sich durch andere Indikatoren ersetzen, das heißt die Indikatoren sind austauschbar. • Die Indikatoren besitzen inhaltlich denselben Kern, messen also gewissermaßen dasselbe.	• Bei Elimination eines Indikators verändert sich die inhaltliche Aussage, die hinter der LV steht.
Kovariation von Indikatoren und LV	• Ändert einer von mehreren gleich kodierten Indikatoren plötzlich seine Ausprägung in eine bestimmte Richtung, verändern sich alle übrigen Indikatoren notwendigerweise in gleicher Weise. • Eine Veränderung der LV geht notwendigerweise mit einer Veränderung aller Indikatoren einher.	• Es sind Konstellationen denkbar, in denen die Indikatoren untereinander nicht hoch korrelieren. • Die Veränderung eines Indikators geht notwendigerweise mit einer Veränderung der LV einher.

Christophersen et al. [90] und Eberl [91] zeigten in den Wirtschafts- und Sozialwissenschaften eine Dominanz des reflektiven Messmodells auf und führten diesbezüglich verschiedene Gründe

an. Als wesentliche Ursache für die Häufigkeit an Fehlspezifikationen führten sie die ungenügende Verbreitung um das Wissen über den Unterschied zwischen reflektiven und formativen Modellen an. Des Weiteren spielten Akzeptanzprobleme auf Seiten der Forscher sowie die geläufigere Anwendung statistischer Verfahren wie der Faktorenanalyse, die jedoch zur Untersuchung formativer Zusammenhänge ungeeignet ist, eine Rolle in der Präferenz reflektiver Messmodelle [90, 91].

2.6. Fragestellung der Arbeit

Die vorliegende Arbeit untersucht die Frage, ob die von Breckwoldt et al. empirisch basierten zehn Indikatoren in der Lage sind, Unterrichtsqualität aus studentischer Perspektive in der universitären medizinischen Ausbildung valide zu erfassen. Wie in Abschnitt 2.2. dargelegt, beruhen akademische Lehrevaluationen großteilig auf studentischen Beurteilungen. Daher soll in dieser Arbeit explizit der Blickwinkel von Studierenden beleuchtet werden. Es soll dabei untersucht werden, wie groß der jeweilige Einfluss der zehn Kriterien auf die studentische Bewertung von Unterrichtsqualität ist und welche Indikatoren dabei den größten Stellenwert in der Beurteilung einnehmen. Des Weiteren soll geklärt werden, ob der Einfluss der Kriterien auf die Bewertung von Unterricht sich in verschiedenen Unterrichtsformaten oder zu verschiedenen Zeitpunkten im Studienverlauf unterscheidet.

Zur Untersuchung dieser Fragestellungen wurden die zehn Kriterien in ein Messmodell als Indikatoren der latenten Variabel „Unterrichtsqualität“ aufgenommen. Um die Beziehung zwischen der latenten Variabel und den Indikatoren zu definieren, wurden die Entscheidungskriterien aus Abschnitt 2.5.3. angewendet. Die zehn Kriterien stellen die Ursache der latenten Variabel Unterrichtsqualität dar, sie sind nicht untereinander austauschbar, sondern bestimmen in ihrer Kombination den Inhalt der latenten Variabel. Somit kann das theoretische Konstrukt Unterrichtsqualität als gewichtete Zusammensetzung seiner Indikatoren, der zehn Kriterien, betrachtet werden und es handelt sich um formative Indikatoren. Um das formative Messmodell zu schätzen und in seiner Güte zu beurteilen und damit die Frage zu beantworten, ob Unterrichtsqualität durch die zehn Kriterien valide gemessen werden kann, wurde ein erweitertes Strukturmodell entwickelt. Anhand der Empfehlungen zur Gütebeurteilung von formativen Messmodellen aus Abschnitt 2.5.2.1. wurde zusätzlich die Globalbewertung der Unterrichtsqualität in das Modell mit aufgenommen. Diese kann als eine direkte Beobachtung des theoretischen Konstrukts Unterrichtsqualität angenommen werden und dient somit als abhängige Variabel für die zu validierenden Indikatoren (Abbildung 4).

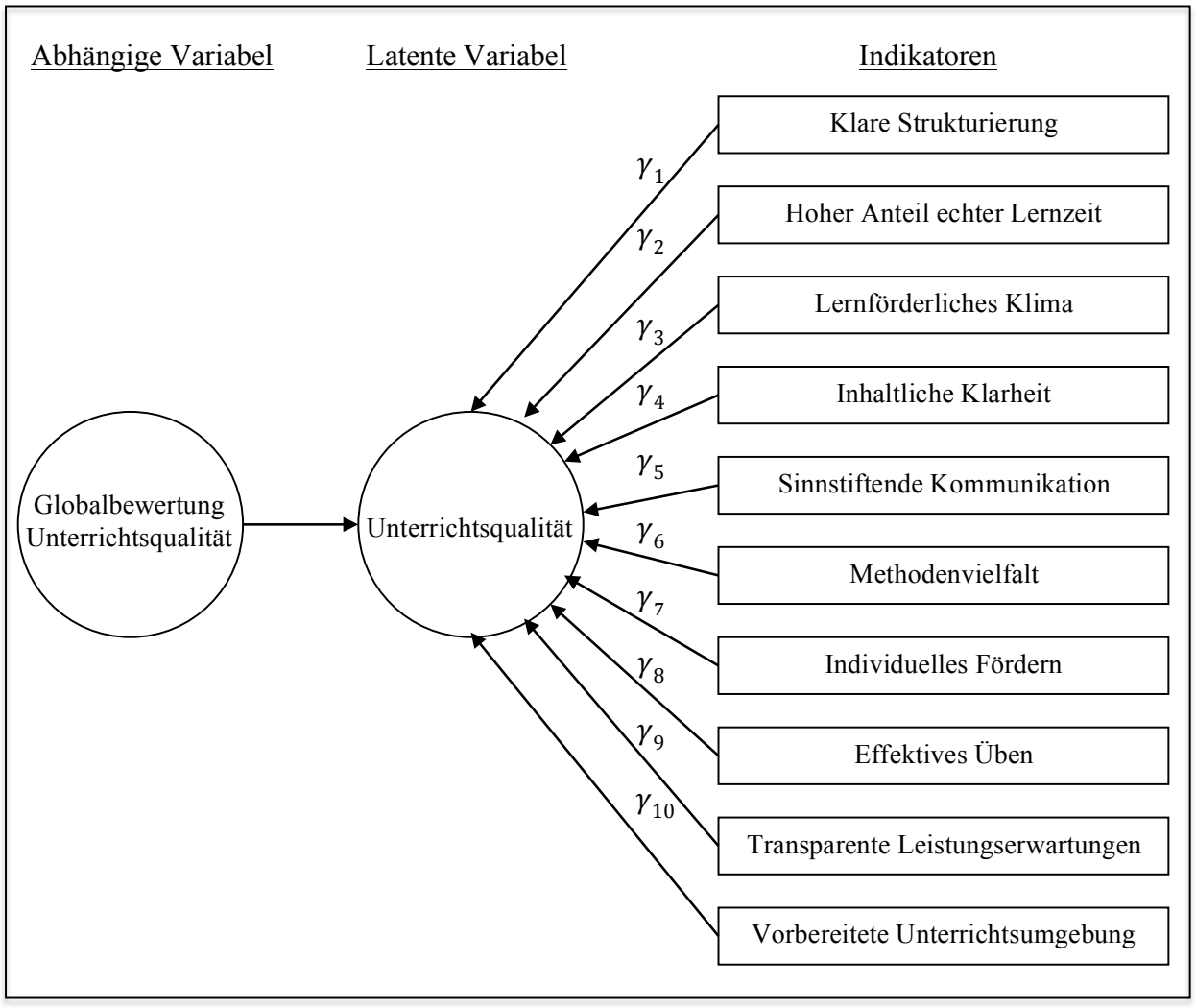

Abbildung 4: Erweitertes formatives Messmodell zur Operationalisierung des theoretischen Konstrukts Unterrichtsqualität durch die zehn empirisch basierten Kriterien $\boldsymbol{\gamma}$: Gewicht

3. METHODIK

3.1. Berlin Teaching Quality Questionnaire 10 (BTQ-10)

Zur Bewertung der einzelnen zehn Kriterien sowie der Globalbewertung der Unterrichtsqualität wurde der Berlin Teaching Quality Questionnaire 10 (BTQ-10) entwickelt (siehe Abbildung 5). Jedes der zehn Kriterien wurde durch typische beobachtbare Beispiele spezifiziert, um zu einem einheitlichen Verständnis der eher abstrakten Kriterien zu gelangen. Die jeweiligen Beispiele stützen sich auf die Beschreibungen jedes Kriteriums, wie sie in Abschnitt 2.2. vorgenommen wurden. Dabei wurden nur positiv gepolte Aussagen benutzt. Zur quantitativen Beurteilung wurde eine fünfstufige bipolare numerische Ratingskala von +2 bis -2 gewählt [76, 77, 108]. Die Endpunkte wurden mit verbalen Ankern belegt (+2: „trifft 100% zu“, -2: „trifft überhaupt nicht zu“).

3.2. Unterrichtsvisitationen

Im Sommersemester 2008 wurden 28 Lehrveranstaltungen im klinischen Abschnitt des Regelstudiengangs Humanmedizin an der Charité Universitätsmedizin Berlin (Campus Charité Mitte und Campus Virchow Klinikum) mit dem Instrument BTQ-10 bewertet. Vierzehn davon waren Seminare mit bis zu 20 Studierenden und weitere vierzehn waren „Unterricht am Krankenbett“ mit bis zu sechs Studierenden. Je sieben Veranstaltungen der beiden Unterrichtsformate lagen im 3. klinischen Semester (7. Fachsemester) bzw. im 5. klinischen Semester (9. Fachsemester) in den Fachdisziplinen Urologie, Pädiatrie und Radiologie (siehe Tabelle 3).

Tabelle 3: Zusammensetzung der ausgewählten Lehrveranstaltungen

Unterrichtsform / Fachsemester	Unterricht am Krankenbett		Seminar		
	Anzahl	Studienfach	Anzahl	Studienfach	Gesamt
7. Fachsemester	7	Urologie	7	Radiologie	14
9. Fachsemester	7	Pädiatrie	7	Radiologie	14
Gesamt	14		14		28

CHARITÉ CAMPUS BENJAMIN FRANKLIN

Berlin Teaching Quality Questionnaire 10
Sommersemester 2008

Klinik für Anästhesiologie und operative Intensivmedizin

Fragebogen für Studierende

Im folgenden Fragebogen bitten wir Sie, die Aussagen zu bewerten:
Die Bewertung bezieht sich auf den Gesamtpunkt, die einzelnen Unterpunkte sollen Ihnen als Hilfestellung dienen. Bitte geben Sie pro Kästchen nur eine Bewertung ab.

+2 bedeutet, das Kriterium trifft 100% zu, -2 heißt, es trifft überhaupt nicht zu.

Herzlichen Dank, dass Sie sich die Zeit nehmen, den Bogen gewissenhaft auszufüllen!

	trifft 100% zu				trifft überhaupt nicht zu
	+2	+1	0	-1	-2
Klare Strukturierung → Es gab einen „roten Faden" während des Unterrichts. → Die Lernschritte bauten thematisch aufeinander auf. → Die Aufgaben waren klar formuliert. → Der Lernstoff wurde zunächst erarbeitet, geübt oder verfestigt *und* am Ende wurden die Ergebnisse, z.B. in Form von Zusammenfassungen, Wiederholungen etc., gesichert.	○ +2	○ +1	○ 0	○ -1	○ -2
Hoher Anteil echter Lernzeit → Der Unterricht hat pünktlich begonnen und geendet. → Der Dozent war organisatorisch und thematisch gut vorbereitet. → Das Lerntempo während des Unterrichts war angemessen. → Organisatorisches wurde aus dem Unterricht ausgelagert oder möglichst kurz gehalten.	○ +2	○ +1	○ 0	○ -1	○ -2
Lernförderliches Klima → Es herrschte gegenseitiger Respekt und die einzelnen Studierenden wurden vom Dozenten gleich behandelt. → Es wurden Regeln, z.B. keine Benutzung von Mobiltelefonen etc., eingehalten. → Es herrschte innerhalb der Gruppe eine freundliche Atmosphäre. → Der Dozent unterstützte die aktive Beteiligung aller Studierenden.	○ +2	○ +1	○ 0	○ -1	○ -2
Inhaltliche Klarheit → Die gestellten Aufgaben waren inhaltlich gut verständlich → Der Dozent hat das bereits vorhandene Wissen der Studierenden in Erfahrung gebracht und an dieses inhaltlich angeknüpft. → Die Ergebnisse des Unterrichts wurden klar und verbindlich festgehalten, z.B. mit Zusammenfassungen, Skripten etc. → Der Dozent hat das Thema so aufbereitet, dass die Studierenden jederzeit folgen konnten. Er hat bemerkt und reagiert, wenn jemand nicht folgen konnte.	○ +2	○ +1	○ 0	○ -1	○ -2
Sinnstiftende Kommunikation → Der Nutzen des Kursinhalts für meine weitere Ausbildung, berufliche Zukunft oder auch darüber hinaus war mir klar. → Dieser wurde gefördert durch Interaktionen, Feedback etc. → Das Feedback gestaltete sich sachlich, konstruktiv und fair.	○ +2	○ +1	○ 0	○ -1	○ -2

- bitte wenden -

CHARITÉ CAMPUS BENJAMIN FRANKLIN

	trifft 100% zu				trifft überhaupt nicht zu
Methodenvielfalt → In dieser Unterrichtseinheit kamen verschiedene Lehrmethoden zum Einsatz, z.B. verschiedene Medien, Patienten, Simulationspatienten etc. → Es wurden verschiedene Unterrichtsformen genutzt, z.B. Rollenspiele, Diskussionen, Einzel- oder Gruppenübungen, Vorträge etc. → Die verschiedenen Methoden hatten eine sinnvolle Abfolge.	○ +2	○ +1	○ 0	○ -1	○ -2
Individuelles Fördern → Es wurden alle Studierende in den Unterricht miteinbezogen. → Dem Dozenten ist aufgefallen, wenn einzelne Studierende Hilfestellungen benötigten und förderte sie durch geeignete Massnahmen. → Der Dozent gab den Studierenden eine individuelle und zeitnahe Rückmeldung über ihren Lernfortschritt → Es wurden individuelle Fördermöglichkeiten angeboten, z.B. OP-Hospitationen, E-Mail-Kontakt mit dem Dozenten etc.	○ +2	○ +1	○ 0	○ -1	○ -2
Intelligentes Üben → In dieser Unterrichtseinheit wurde ausreichend oft und sinnvoll geübt. (ACHTUNG: Üben muss nicht nur praktisch sondern kann auch theoretisch sein, z.B. Anwendung des Erarbeiteten auf ein Fallbeispiel, eine Aufgabe, im Rollenspiel etc.) → Die Übungsaufgaben passten zum Lernstand und waren in der vorgegebenen Zeit zu bewältigen. → Der Dozent gab gezielte Hilfestellungen beim Üben.	○ +2	○ +1	○ 0	○ -1	○ -2
Transparente Leistungserwartungen → Mir waren die Lernziele bekannt und verständlich. → Mir wurden zeitnah Rückmeldungen zum Lernfortschritt gegeben. → Durch den Kurs habe ich einen Anreiz erhalten mich weiterführend mit dem Thema zu beschäftigen und meine Lernmotivation wurde gesteigert.	○ +2	○ +1	○ 0	○ -1	○ -2
Vorbereitete Unterrichtsumgebung → Es stand ein den Umständen, z.B. der Gruppengrösse etc., entsprechender und vorbeiteter Unterrichtsraum zur Verfügung. → Das benötigte Unterrichtsmaterial war vorhanden und funktionsfähig. → Der Dozent war mit den Unterrichtsmaterialen, z.B. Gerätetechnik etc., vertraut.	○ +2	○ +1	○ 0	○ -1	○ -2
Insgesamt befand ich den Unterricht für ausgesprochen gut.	○ +2	○ +1	○ 0	○ -1	○ -2

Freitext für Bemerkungen, Anregungen sowie Kritikpunkte:

Vielen Dank!

Abbildung 5: Berlin Teaching Quality Questionnaire 10 (BTQ-10)

3.3. Datenerhebung - und sicherheit

Ein geschulter unabhängiger Beobachter nahm an den ausgewählten Lehrveranstaltungen teil. Vor Unterrichtsbeginn wurden die Dozierenden und Studierenden über die Ziele der Untersuchung informiert und zur Teilnahme eingeladen. Es wurde allen zugesichert, dass die Ablehnung der Studienteilnahme keine negativen Konsequenzen bezüglich der Kursbeteiligung, Benotung oder Personalbewertung nach sich ziehe. Der Beobachter war weder in den Unterricht noch in die Prüfung der Studierenden involviert.

Die Studierenden erhielten den Fragebogen vor Unterrichtsbeginn ausgehändigt. Rückfragen zum BTQ-10 wurden bei Bedarf direkt beantwortet. Die Studierenden wurden gebeten, den BTQ-10 am Ende der Lehrveranstaltung gewissenhaft auszufüllen und an den Beobachter zu retournieren. Personenspezifische Daten der Studierenden wurden nicht erhoben.

Die Lehrkoordinatoren der einzelnen Studienfächer sowie die Ausbildungskomission der Charité - Universitätsmedizin Berlin stimmten den Unterrichtsvisitationen zu. Die Ethikkomission der Charité - Universitätsmedizin Berlin stimmte dem Studienvorhaben zu (Antragsnummer EA 2/009/16).

3.4. Statistische Datenauswertung

Die statistische Auswertung erfolgte mit dem Software Programm „SPSS für Windows“, Version 19.0 (SPSS Inc., Chicago, IL) sowie Stata Version 12 (StataCorp. 2011. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP).

Wie in Abschnitt 2.6. dargestellt, wurden die zehn Kriterien als formative Indikatoren in ein erweitertes Strukturmodell eingebunden. Um den Einfluss der Kriterien auf die Globalbewertung von Unterrichtsqualität zu untersuchen, wurde eine ordinale Regressionsanalyse mit der Globalbewertung als zu erklärende abhängige Variable und den zehn Indikatoren als unabhängige Variablen durchgeführt.

Die Zuordnung von Ratingskalen zur Ordinal- oder Intervallskala ist umstritten und die Hypothese der Intervallskalenqualität von Ratingskalen muss in jeder Untersuchung neu begründet werden [76, 109-117]. Entscheidender als die Bestimmung des Skalenniveaus aus messtheoretischer Sicht ist die Überprüfung der mathematisch-statistischen Voraussetzungen zur Anwendung von parametrischen Tests [76]. Dazu wurde der Datensatz auf Normalverteilung untersucht.

3.4.1. Fallzahlkalkulation

Eine Mindestfallzahl für die Durchführung einer ordinalen Regressionsanalyse findet sich in der Literatur nicht [118]. Es liegen jedoch verschiedene Empfehlungen vor. So sollte die Anzahl der Beobachtungen größer als die mit 5 multiplizierte Anzahl aller Kategorienstufen sein [119]. Die in dieser Arbeit angewandten Regressionsanalysen haben 5 Kategorienstufen, damit sollten mindestens 25 Beobachtungen vorliegen. Eine weitere Daumenregel empfiehlt ein Zehnfaches der Anzahl der einbezogenen Kovariaten [119, 120]. Das heißt, dass bei zehn Kovariaten mindestens 100 Fälle vorliegen sollten. Restriktivere Empfehlungen besagen, dass die Fallzahl bei n>100 und die Anzahl der Beobachtungsfälle einer Kategorie der abhängigen Variabel bei n>25 liegen sollten [120, 121,122]. In der vorliegenden Arbeit sind diese Empfehlungen erfüllt.

3.4.2. Ordinale Regressionsanalyse

Mit Hilfe von Regressionsanalysen können Beziehungen zwischen einer abhängigen zu erklärenden Variablen und einer oder mehreren unabhängigen erklärenden Variablen modelliert werden. Während bei der linearen Regression von einem linearen Zusammenhang ausgegangen wird und ein metrisches Messniveau der abhängigen Variablen sowie normalverteilte und varianzhomogene Residuen gefordert werden, können bei nominal- oder ordinalskalierten abhängigen Variablen logistische Regressionsmodelle angewendet werden [122-127].

Im Gegensatz zur linearen Regression versucht die logistische Regression dabei nicht Schätzungen für die Beobachtungen der abhängigen Variablen vorzunehmen, sondern die Eintrittswahrscheinlichkeiten dieser Beobachtungswerte aufgrund der Ausprägung einer oder mehrerer unabhängiger Variablen abzuleiten. Die unabhängigen Variablen können dabei jedes beliebige Skalenniveau aufweisen.

Im binären Fall verfügt die abhängige Variable Y über die Ausprägungen 1 und 0, wobei y=1 als „Ereignis tritt ein“ und y=0 als „Ereignis tritt nicht ein“ interpretiert wird. Die Koeffizienten $ß_j$ des Regressionsmodells, die auch als Logit-Koeffizienten bezeichnet werden, spiegeln die Einflussstärke der unabhängigen Variablen X_j auf die Eintrittswahrscheinlichkeit des Ereignisses wider. Eine einfache Interpretation dieser Logit-Koeffizienten, wie etwa bei der linearen Regressionsanalyse, ist jedoch nicht möglich, da der Zusammenhang zwischen den unabhängigen und der abhängigen Variablen nicht linear ist. Lediglich die Richtung des Zusammenhanges kann angegeben werden. Negative Koeffizienten führen bei steigenden x-Werten zu einer kleineren relativen Wahrscheinlichkeit und positive Koeffizienten zu einer größeren relativen Wahrscheinlichkeit des Eintretens des Ereignisses. Die Koeffizienten eines

Modells sind untereinander jedoch nicht vergleichbar. Zur Erleichterung der Interpretation wird deshalb das Verhältnis der Eintrittswahrscheinlichkeit P(y=1) zur Gegenwahrscheinlichkeit P(y=0) betrachtet. Dieses Wahrscheinlichkeitsverhältnis spiegelt die Chance (Odd) wider, das Ereignis y=1 im Vergleich zum Ereignis y=0 zu erhalten. Es werden die Effekt-Koeffizienten e^{β} berechnet, die das Verhältnis angeben, in dem die Odds vor und nach einer Veränderung von x um eine Einheit zueinanderstehen. Diese spiegeln somit den Faktor wider, um den sich das Wahrscheinlichkeitsverhältnis ändert, wenn eine unabhängige Variabel um eine Einheit erhöht wird und alle anderen unverändert bleiben. Sie können als Odds-Ratios interpretiert werden. Sie können nur Werte zwischen 0 und $+\infty$ annehmen. Ein $e^{\beta} < 1$ spricht für einen negativen Zusammenhang, ein $e^{\beta} > 1$ für einen positiven Zusammenhang und der Wert 1 für keinen Zusammenhang zwischen der abhängigen und den unabhängigen Variablen. Im Gegensatz zu den Logit-Koeffizienten können die Effekt-Koeffizienten bzw. Odds-Ratios der unabhängigen Variablen innerhalb eines Modells untereinander verglichen werden [126, 128].

Um die Gültigkeit des ordinalen Regressionsmodells zu bestätigen, wurde die „Annahme paralleler Regressionen“ (proportional odds assumption) überprüft. Dabei handelt es sich um eine Grundvoraussetzung zur Durchführung einer ordinalen logistischen Regression. In dem Modell der ordinalen logistischen Regression wird versucht, die Wahrscheinlichkeit für das Auftreten der Kategorien K (k=1, 2, 3, …K) einer ordinal skalierten abhängigen Variabel durch die unabhängigen Variablen vorherzusagen. Zur Berechnung findet das Modell der kumulierten Logits (cumulative logit model, proportional odds model) Anwendung. Dabei wird für die Logit-Koeffizienten angenommen, dass sie von den Kategorien K der abhängigen Variabel unabhängig sind. Das heißt, dass der Anstieg einer unabhängigen Variablen um eine Einheit den gleichen Effekt hat, unabhängig davon, ob die abhängige Variable zwischen der ersten und zweiten Kategorie oder zwischen den Kategorien K-1 und K wechselt. Somit lässt sich der Einfluss der unabhängigen Variablen jeweils durch einen universellen Koeffizienten beschreiben, der für jeden Stufenwechsel innerhalb des ordinalen Modells gültig ist. Um diese „Annahme paralleler Regressionen“ zu bestätigen, wird ein Regressionsmodell mit und eines ohne Annahme der parallelen Regressionen berechnet. Mittels Chi2-Test wird überprüft ob die -2LogLikelihood-Werte beider Modelle annähernd gleich sind. Die Nullhypothese besagt, dass parallele Regressionen vorliegen. Wird die Nullhypothese verworfen, verliert das Modell seine Gültigkeit [129].

3.4.3. Multikollinearität

Es wurde überprüft, ob unter den zehn formativen Indikatoren Multikollinearität besteht, die in dem ordinalen Regressionsmodell berücksichtigt werden muss.

Multikollinearität bezeichnet eine lineare Abhängigkeit zwischen zwei oder mehreren unabhängigen Variablen [101, 130]. Lässt sich eine unabhängige Variabel als lineare Funktion einer oder mehrerer anderer unabhängiger Variablen darstellen, spricht man von perfekter Multikollinearität. Dabei ist die Berechnung der Regressionskoeffizienten nicht möglich [131]. Perfekte Multikollinearität tritt meist dann auf, wenn dieselbe Einflussgröße zweimal als unabhängige Variabel in das Regressionsmodell aufgenommen wird. Versteckte hohe Multikollinearität führt zu Verzerrungen der Schätzung der Regressionsparameter und der Standardfehler [122]. Es existieren mehrere Methoden um Multikollinearität nachzuweisen [101]. Zum Ausschluss von Multikollinearität zwischen den Indikatoren wurden die zwei gebräuchlichsten Kriterien angewandt und die Toleranz (Tol) bzw. der Variance-Inflation-Factor (VIF) sowie die Konditionsindices betrachtet.

Anhand einer Hilfsregression jeder unabhängigen Variabel auf die übrigen unabhängigen Variablen, lässt sich nachweisen, ob sich eine unabhängige Variabel durch Linearkombination der anderen darstellen lässt. Aus dem resultierenden R^2 lässt sich dann die Toleranz und der Variance-Inflation-Factor berechnen, wobei gilt:

$$VIF = \frac{1}{Tol} = \frac{1}{1 - R^2}.$$

VIF-Werte von über 10 geben einen sicheren Hinweis auf Multikollinearität [101, 130]. Ein Konditionsindex, der aus den Eigenwerten der Schätzung ermittelt wird, von 30-100 spricht für starke Multikollinearität. Durch eine Varianz-Zerlegungsmatrix der unabhängigen Variablen, die für jede unabhängige Variable den Varianzanteil zeigt, den jeder Konditionsindex verursacht, können die unabhängigen Variablen identifiziert werden, für die der Verdacht auf Multikollinearität besteht. Dazu wird für jeden kritischen Konditionsindex untersucht, ob für zwei oder mehrere unabhängige Variablen ein Varianzanteil von größer 0,5 auf diesen Konditionsindex zurückzuführen ist [101].

Beim Nachweis von Multikollinearität werden die untereinander multikollinearen Indikatoren zu einem Index zusammengefasst, der im Rahmen des Messmodells wie ein einzelner Indikator behandelt wird [90, 101, 103]. Die Gewichtung der Indikatoren richtet sich nach der Art der Beziehung der Indikatoren untereinander. Stehen sie in einem kompensatorischen Verhältnis,

wird das arithmetische Mittel der Indikatoren gebildet [101, 103]. Auf Basis einer explorativen Faktorenanalyse ist auch eine Bündelung von Indikatoren möglich. Eine Elimination multikollinearer Indikatoren in formativen Messmodellen sollte nicht erfolgen, da sie mit dem Verlust einer relevanten inhaltlichen Facette des Konstrukts verbunden ist [87, 89, 90, 92, 94, 103].

3.4.4. Schätzung und Gütebeurteilung

Die Schätzung der logistischen Regressionsfunktion erfolgt über die Maximum-Likelihood-Methode. Ziel des Verfahrens ist es, die Logit-Koeffizienten so zu bestimmen, dass die Wahrscheinlichkeit (Likelihood), die beobachteten Daten zu erhalten, maximiert wird. Die Schätzung erfolgt iterativ mit Hilfe des Newton-Raphson-Algorithmus, bei dem die Logit-Koeffizienten systematisch geändert werden, bis die logarithmierte Likelihood-Funktion (LogLikelihood-Funktion) sich nicht mehr deutlich vergrößern lässt.

Zur Beurteilung der Güte des Regressionsmodells fanden der Likelihood-Ratio-Test, die Pseudo-R^2-Statistiken sowie die Analyse der Klassifikationsmatrix Anwendung.

Beim Likelihood-Ratio-Test (LR), auch Modell-Chi-Quadrat-Test, wird zunächst ein Nullmodell berechnet, in das nur die Konstante einfließt und alle Regressionskoeffizienten gleich Null sind. Im nächsten Schritt wird das vollständige Modell unter Berücksichtigung aller erklärenden Variablen berechnet. Die absolute Differenz zwischen den -2fachen Log-Likelihood-Werten beider Modelle, auch Devianz genannt, wird anhand des Chi²-Tests auf Signifikanz überprüft.

Die Pseudo-R^2-Statistiken streben eine Interpretation analog zum R^2 der linearen Regression als Maß der Erklärungskraft des Modells an. Das Bestimmtheitsmaß R^2 der linearen Regression gibt den Anteil der durch das Regressionsmodell erklärten Varianz an der Gesamtvarianz der abhängigen Variabel wider. Da die Varianz des Modells in der logistischen Regression nicht direkt berechnet werden kann, beruhen die Pseudo- R^2-Statistiken auf dem Verhältnis zwischen dem Likelihood bzw. Log-Likelihood des Nullmodells und dem des vollständigen Modells. Die drei gebräuchlichsten Maßzahlen sind die Pseudo-R^2 Maße nach McFadden (R^2_{McF}), Cox und Snell (R^2_{CS}) und Nagelkerke (R^2_N) [132,133,134]. Bei Werten über 0,2 kann das Modell als akzeptabel, bei Werten über 0,4 als gut eingestuft werden. Werte über 0,5 für Nagelkerke-R^2 gelten als sehr gut [121, 122, 128].

Bei der Analyse der Klassifikationsmatrix werden die beobachten mit den modellierten Gruppenzuordnungen verglichen. Der Anteil der korrekt zugeteilten Fälle spiegelt die Güte des

Modells wieder und sollte höher sein als eine rein zufällige Zuordnung, die durch die proportionale Zufallswahrscheinlichkeit (PZW; proportional chance criterium, PCC) beschrieben werden kann. Im ordinalen Fall berechnet sich die proportionale Zufallswahrscheinlichkeit wie folgt:

$$\text{PZW} = \sum_{g=1}^{G} \left(\frac{n_g}{n}\right)^2 = \sum_{g=1}^{G} a_g^2$$

mit n_g: Anteil der Elemente in Gruppe g (g=1,…,G),

n: Gesamtstichprobe,

a_g: Anteilswert der Gruppe g (g=1,…, G) an der Gesamtstichprobe n,

G: Anzahl der Gruppen.

Tabelle 4 gibt eine Zusammenfassung der Gütemaße und deren Wertebereiche wieder.

Die einzelnen Koeffizienten der logistischen Regressionsanalyse wurden mittels der Wald-Statistik auf Signifikanz geprüft. Dabei wird der Quotient des quadrierten Regressionskoeffizienten und dem Standardfehler anhand des Chi²-Tests auf Signifikanz überprüft.

Tabelle 4: Gütemaße der logistischen Regressionsanalyse

Gütemaß	Wertebereich
Likelihood-Ratio-Test	Möglichst hoher Chi^2-Wert; Signifikanzniveau < 5%
McFadden-R^2	>0,2 akzeptabel; >0,4 gut
Cox und Snell R^2	>0,2 akzeptabel; >0,4 gut
Nagelkerke R^2	>0,2 akzeptabel; >0,4 gut; >0,5 sehr gut
Klassifikationsmatrix	Wert der korrekten Klassifikationen > PZW

PZW: proportionale Zufallswahrscheinlichkeit

3.4.5. Heterogeneous Choice Modelle

Um zu untersuchen, ob sich der Einfluss der zehn Kriterien auf die Globalbewertung von Unterrichtsqualität zwischen den beiden Unterrichtsformaten sowie zwischen den beiden gewählten Zeitpunkten im Studienverlauf unterscheidet, wurde das ordinale Regressionsmodell auf signifikante Gruppenunterschiede getestet.

Für den Vergleich zweier Gruppen wurde ein Regressionsmodell für die Gesamtstichprobe berechnet, in das eine Dummy Variable für die Gruppenzugehörigkeit sowie ein Interaktionsterm, der das Produkt jeder unabhängigen Variable und der Dummy Variable darstellt, eingefügt wurden [135, 136]. Bei einer Dummy Variable handelt es sich um eine binäre Variable, die mit „0“ für die Referenzgruppe und mit „1“ für die Vergleichsgruppe kodiert. In diesem Modell stellt somit der Koeffizient für jede unabhängige Variable den Koeffizienten für die Referenzgruppe dar. Der Koeffizient für den Interaktionsterm stellt die Differenz zwischen den Koeffizienten für die Referenz- und die Vergleichsgruppe dar. Der p-Wert des Interaktionsterms liefert einen Signifikanztest für die Differenz der Koeffizienten zwischen der Referenz- und Vergleichsgruppe.

Zusätzlich zum ordinalen Regressionsmodell wurde ein Heterogeneous choice Modell berechnet, da beim Vergleich von Logit-Koeffizienten sowie Odds-Ratios über Gruppen hinweg gravierende Fehleinschätzungen auftreten können [135-141]. Dies hat folgenden Grund.

In dem Modell der ordinalen logistischen Regression lässt sich die abhängige Variable Y auf eine latente, nicht empirisch beobachtbare metrische Variable Y* zurückführen, für die folgende lineare Regressionsgleichung gilt:

$$Y^* = \beta_0 + \beta_1 \times x + \ldots + \beta_j \times x_j + \varepsilon$$

mit β_0: Konstante,

β_j: Koeffizienten,

ε: Fehlerterm.

Da die Variable Y* nicht messbar ist, ist der unbekannte Mittelwert sowie die unbekannte Varianz des Fehlerterms ε nicht durch Schätzung aus den Daten bestimmbar. Für den Fehlerterm ε wird daher innerhalb der logistischen Regressionsmodelle eine logistische Verteilung mit dem Erwartungswert 0 und der festen Varianz $\frac{\pi^2}{3}$ angenommen.

Die lineare Regressionsgleichung für die latente Variable Y* kann demzufolge wie folgt erweitert werden:

$$Y^* = \beta_0 + \beta_1 \times X_1 + \ldots + \beta_j \times X_j + \delta \times \varepsilon,$$

dabei fungiert der unbekannte Faktor δ als Regulativ bzw. Skalierungsfaktor des Fehlerterms ε und erlaubt der fixen Varianz von ε den wahren unbekannten Wert anzunehmen [135-137, 141]. Somit werden innerhalb des logistischen Regressionsmodells nicht die wahren β-Koeffizienten aus oben genannter Gleichung bestimmt, sondern δ-normierte Koeffizienten β`. Das Regressionsmodell ist damit unterdeterminiert. Es gilt folgender Zusammenhang:

$$\beta' = \frac{\beta}{\delta}.$$

Finden sich somit in zwei Gruppen verschiedene β´-Koeffizienten, kann nicht entschieden werden, ob wirklich verschiedene Wirkungen, sprich verschiedene wahre β-Koeffizienten, vorliegen oder ob unterschiedliche Residualvarianzen verantwortlich sind. Beispielsweise kann ein halb so großer Logit-Koeffizient β´ in Gruppe 1 gegenüber Gruppe 2 bedeuten, dass δ in Gruppe 1 doppelt so hoch ausfällt als in Gruppe 2 und die Regressionskoeffizienten β jedoch eigentlich identisch sind [135-141]. Somit kann es beim Vergleich von Logit-Koeffizienten sowie Odds-Ratios über Gruppen hinweg aufgrund unbeobachteter Heterogenität der Varianzen (Heteroskedastizität) zu Fehlannahmen kommen. Die Anwendung von Heterogeneous Choice Modellen (location-scale models) erlaubt es diese gruppenspezifische Heterogenität zu modellieren [136, 139-141]. Dies ermöglicht es, eine vorliegende Heteroskedastizität aufzudecken und gegebenenfalls zu modellieren und die beschriebene Unsicherheit beim Gruppenvergleich in logistischen Regressionsmodellen zu reduzieren.

Zum Vergleich der Anpassungsgüte des ordinalen Regressionsmodelle und des Heterogeneous choice Modells wurden das „Akaike Informationskriterium“ (AIC) und das „Bayessche Informationskriterium“ (BIC) eingesetzt. Beide Informationskritierien können im Gegensatz zu den Pseudo-R^2-Statistiken zum Vergleich von geschachtelten (nested models) sowie ungeschachtelten (non-nested models) Modellen benutzt werden. Es wird das Modell vorgezogen, bei dem die Informationskriterien möglichst kleine Werte annehmen [142,143].

4. ERGEBNISSE

4.1. Unterrichtsvisitationen

Von 275 ausgegebenen Bewertungsbögen wurden 256 (93,1%) vollständig ausgefüllt retourniert. Die Rücklaufquote im Unterrichtsformat UaK betrug 100%, im Unterrichtsformat Seminar 90% (Tabelle 5).

Tabelle 5: Anzahl der ausgegebenen und retournierten Bewertungsbögen (BTQ-10)

Studienfach		Ausgegebene BTQ-10	Retournierte BTQ-10	
			Anzahl	Prozent
UaK	Urologie	39	39	100%
	Pädiatrie	36	36	100%
Seminar	Radiologie 7. Fachsemester	89	82	92,1%
	Radiologie 9. Fachsemester	111	99	89,2%
Gesamt		275	256	93,1%

Die Gruppenstärke der Unterrichtsvisitationen ist in Tabelle 6 zusammengefasst.

Tabelle 6: Gruppenstärke pro Studienfach und Unterrichtsformat

Studienfach		Gruppenstärke		
		Minimum	Maximum	Mittelwert
UaK	Urologie	5	6	5,6
	Pädiatrie	4	6	5,1
Seminar	Radiologie 7. Fachsemester	5	16	12,7
	Radiologie 9. Fachsemester	13	20	15,9

4.2. Ergebnisse des BTQ-10 in den Unterrichtsvisitationen

In den Auswertungsbögen ergab sich kein Anhalt für Antworttendenzen wie zum Beispiel das Ankreuzen gleicher Zahlenreihen. In Tabelle 7 ist der Median und Mittelwert der einzelnen zehn Kriterien sowie der Globalbewertung der Unterrichtsqualität angegeben. Das Kriterium „Methodenvielfalt“ wurde durchschnittlich am schlechtesten und das Kriterium „Lernförderliches Klima“ durchschnittlich am besten bewertet.

Für alle zehn Kriterien sowie die Gesamtbewertung lag keine Normalverteilung vor. Der Kolmogorov-Smirnov-Test war jeweils hochsignifikant ($p<0,001$) und die Annahme der Normalverteilung musste verworfen werden.

Tabelle 7: Median, Mittelwert und Standardabweichung der Ergebnisse des BTQ-10 in 28 Unterrichtsvisitationen (n=256)

Kriterien	Median	Mittelwert	SD
Klare Strukturierung	4	3,88	1,04
Hoher Anteil echter Lernzeit	4	4,14	1,01
Lernförderliches Klima	5	4,36	0,84
Inhaltliche Klarheit	4	3,94	1,04
Sinnstiftende Kommunikation	4	3,95	0,98
Methodenvielfalt	3	3,24	1,2
Individuelles Fördern	4	3,35	1,21
Effektives Üben	4	3,44	1,18
Transparente Leistungserwartungen	4	3,45	1,09
Vorbereitete Unterrichtsumgebung	4	4,19	0,89
Gesamtbewertung	4	3,9	0,99

SD: Standardabweichung

Für alle Kriterien sowie für die Globalbewertung wurden sämtliche Antwortstufen der Ratingskala verwendet (siehe Abbildung 6-16). Für die Darstellung der prozentualen Verteilung der Antwortstufen in den Abbildungen 6 bis 16 wurde die Skala im Gegensatz zu den verwendeten Fragebögen von -2 bis +2 aufgetragen. Unter der Freitextoption wurden keinerlei Angaben gemacht.

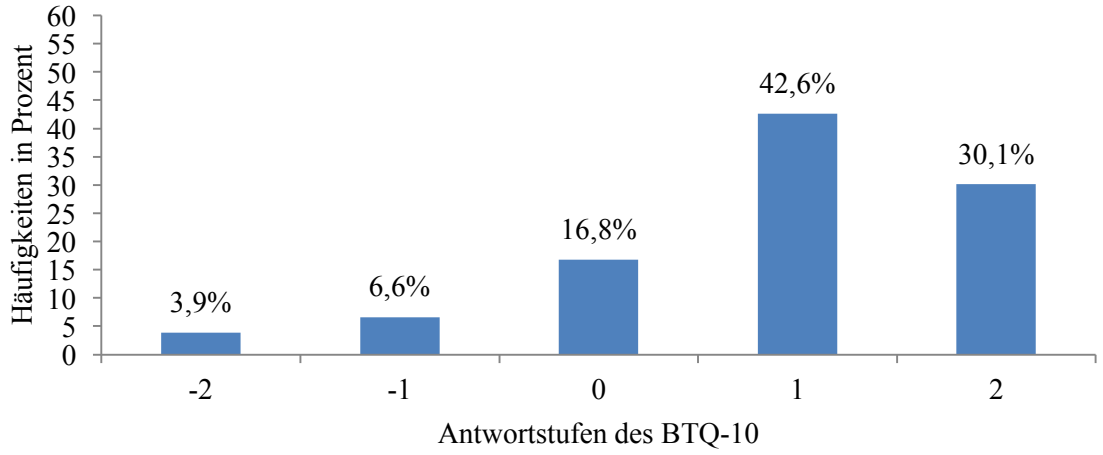

Abbildung 6: Prozentuale Verteilung der Antwortstufen von -2 bis +2 des BTQ-10 für das Kriterium „Klare Strukturierung“ (n=256)

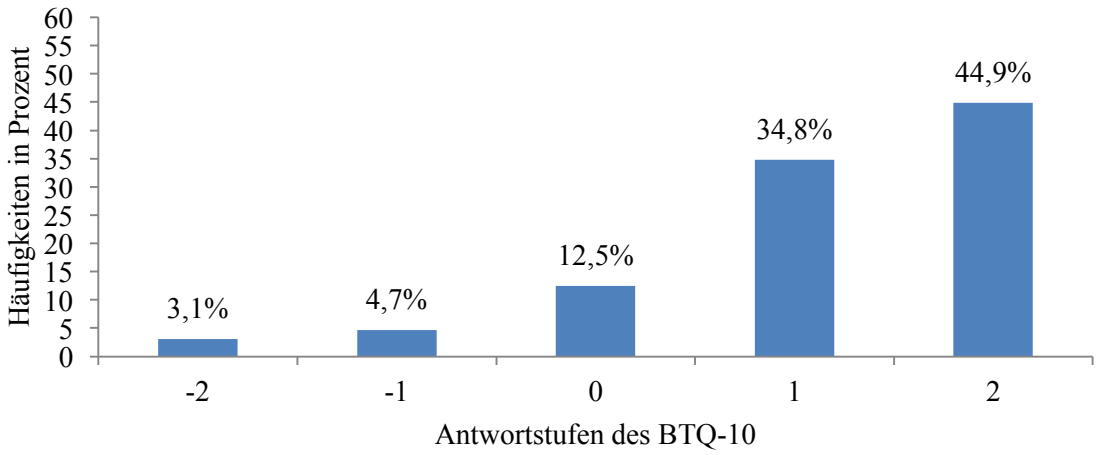

Abbildung 7: Prozentuale Verteilung der Antwortstufen von -2 bis +2 des BTQ-10 für das Kriterium „Hoher Anteil echter Lernzeit“ (n=256)

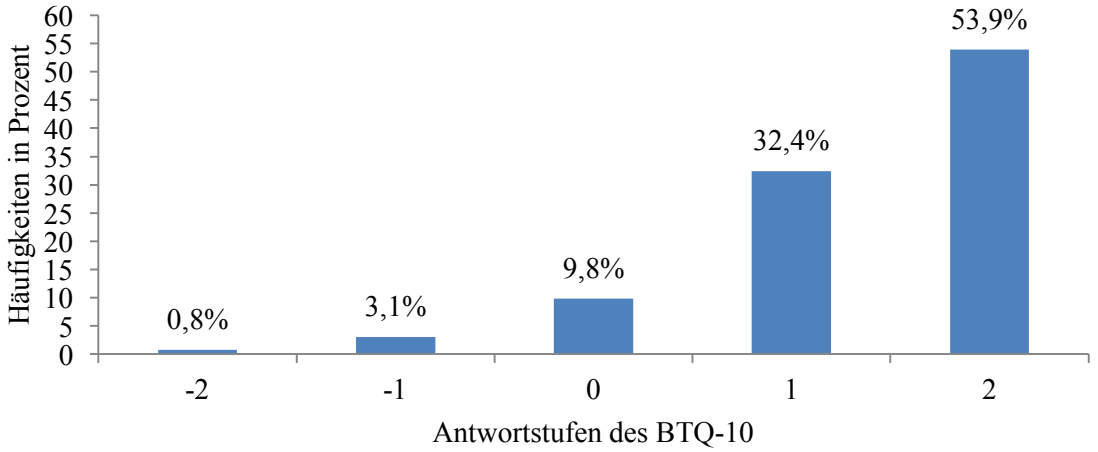

Abbildung 8: Prozentuale Verteilung der Antwortstufen von -2 bis +2 des BTQ-10 für das Kriterium „Lernförderliches Klima“ (n=256)

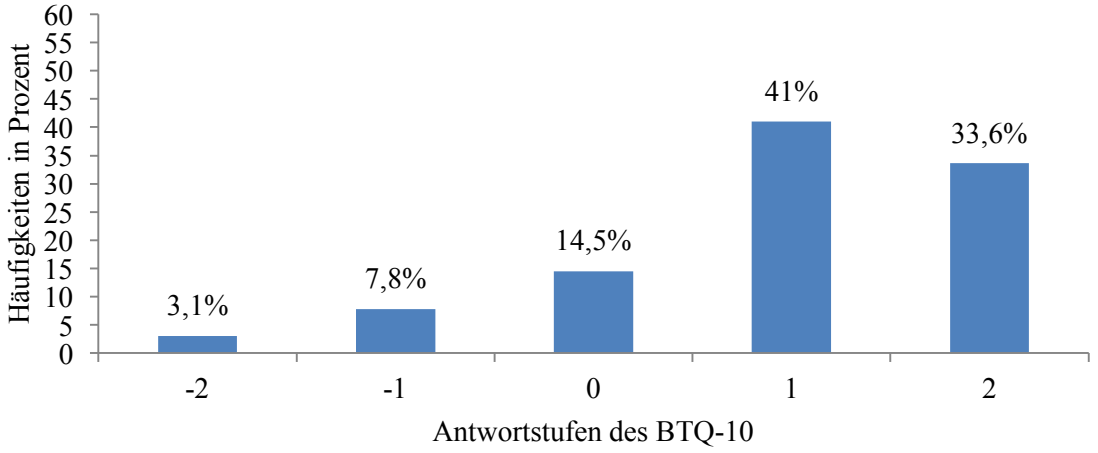

Abbildung 9: Prozentuale Verteilung der Antwortstufen von -2 bis +2 des BTQ-10 für das Kriterium „Inhaltliche Klarheit“ (n=256)

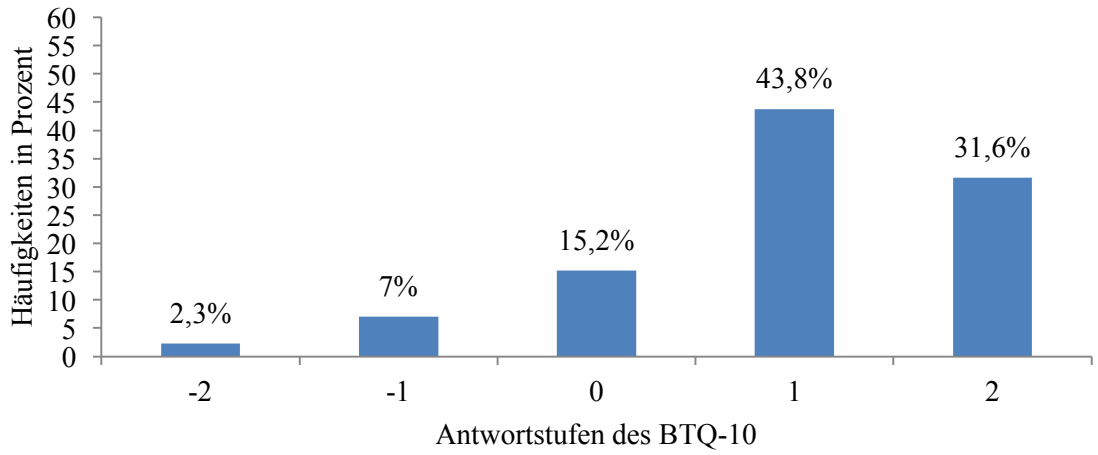

Abbildung 10: Prozentuale Verteilung der Antwortstufen von -2 bis +2 des BTQ-10 für das Kriterium „Sinnstiftende Kommunikation“ (n=256)

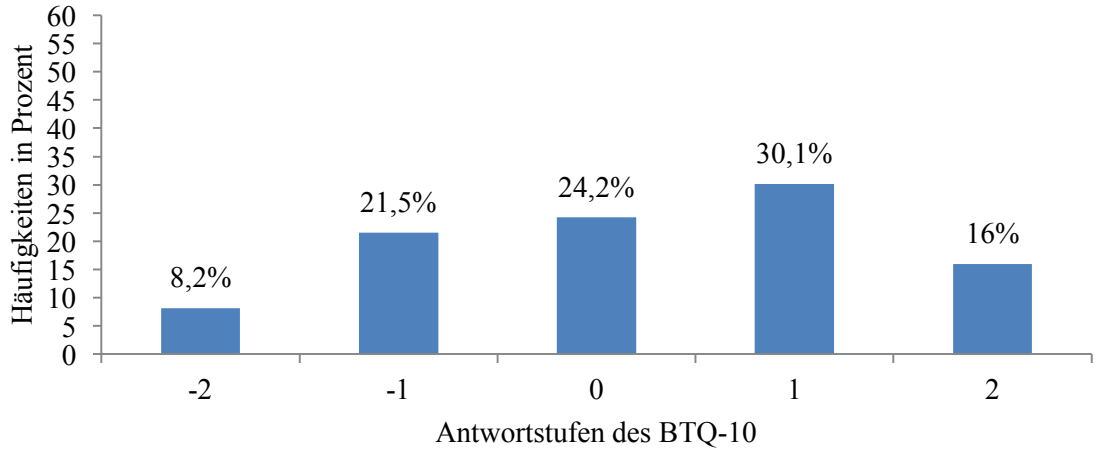

Abbildung 11: Prozentuale Verteilung der Antwortstufen von -2 bis +2 des BTQ-10 für das Kriterium „Methodenvielfalt“ (n=256)

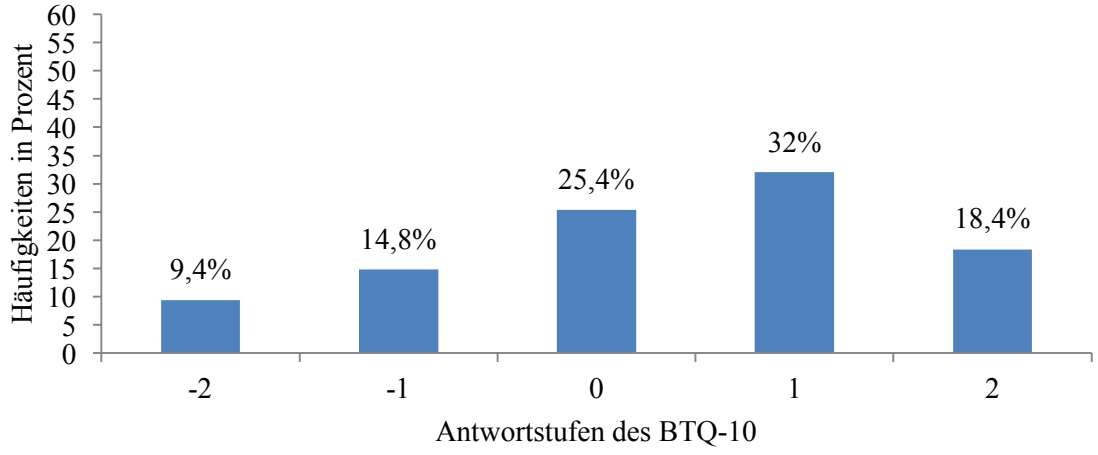

Abbildung 12: Prozentuale Verteilung der Antwortstufen von -2 bis +2 des BTQ-10 für das Kriterium „Individuelles Fördern“ (n=256)

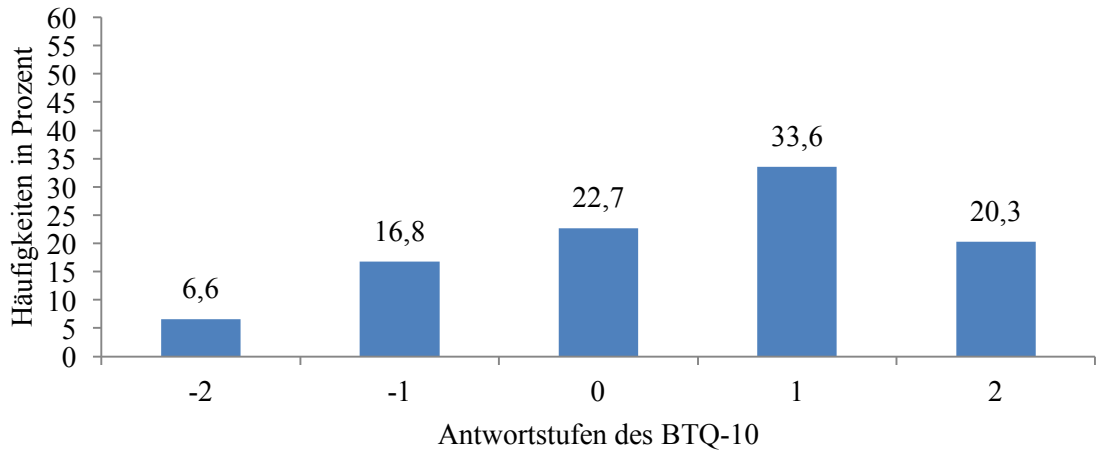

Abbildung 13: Prozentuale Verteilung der Antwortstufen von -2 bis +2 des BTQ-10 für das Kriterium „Effektives Üben“ (n=256)

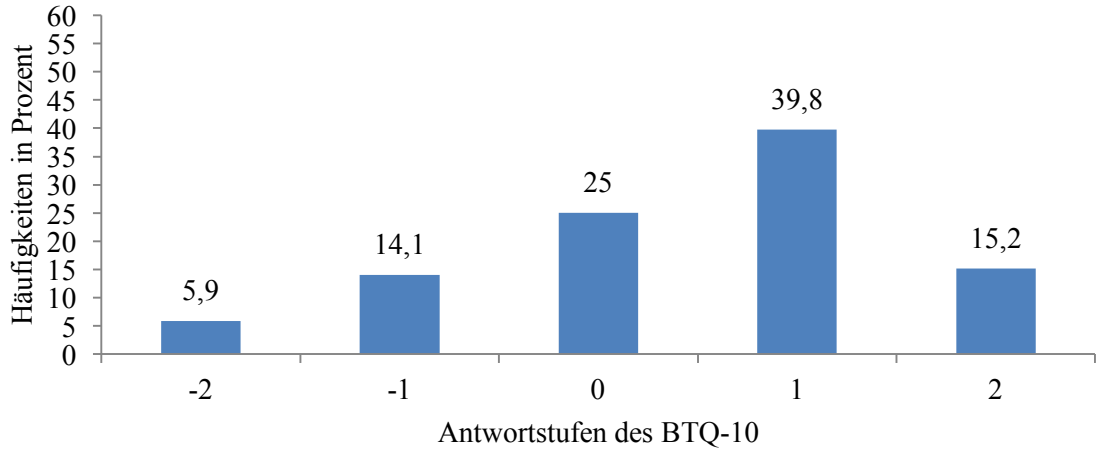

Abbildung 14: Prozentuale Verteilung der Antwortstufen von -2 bis +2 des BTQ-10 für das Kriterium „Transparente Leistungserwartungen“ (n=256)

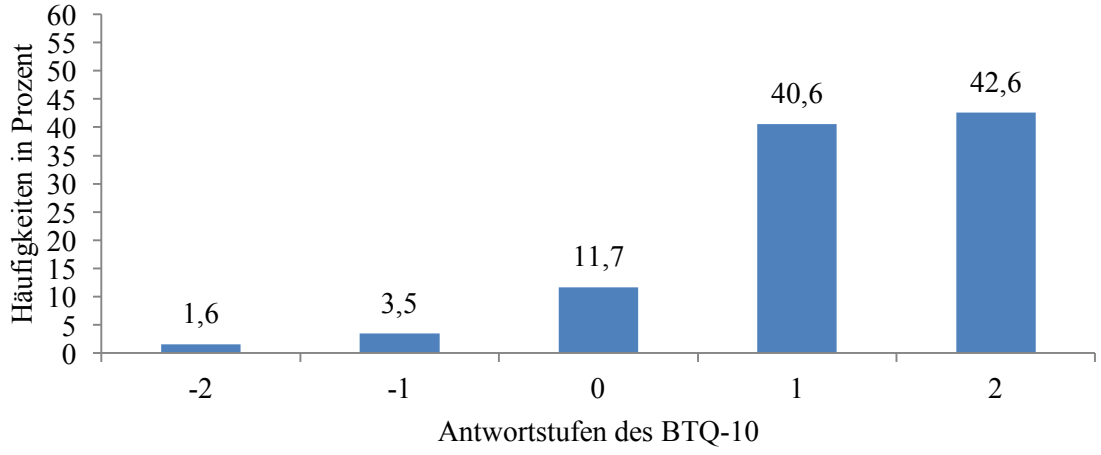

Abbildung 15: Prozentuale Verteilung der Antwortstufen von -2 bis +2 des BTQ-10 für das Kriterium „Vorbereitete Unterrichtsumgebung“ (n=256)

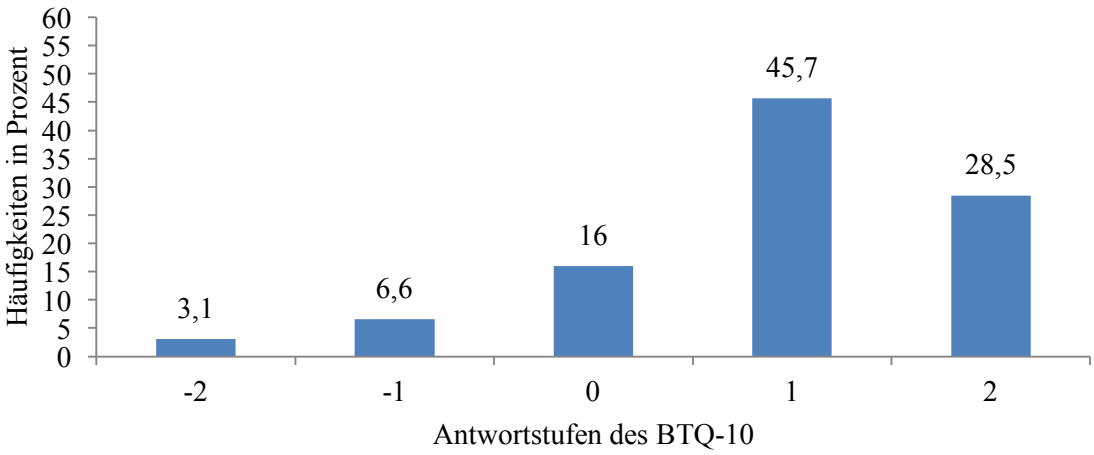

Abbildung 16: Prozentuale Verteilung der Antwortstufen von -2 bis +2 des BTQ-10 für die Globalbewertung (n=256)

4.3. Einfluss der zehn Kriterien auf die Bewertung von Unterrichtsqualität

In der ordinalen Regressionsanalyse (Übersicht siehe Tabelle 9) zur Untersuchung des kausalen Zusammenhanges zwischen den zehn Kriterien und der Globalbewertung von Unterrichtsqualität konnten sieben Kriterien als signifikant positive Einflussfaktoren identifiziert werden:

- Sinnstiftende Kommunikation,
- Inhaltliche Klarheit,
- Klare Strukturierung,
- Individuelles Fördern,
- Hoher Anteil echter Lernzeit
- Transparente Leistungserwartungen,
- Effektives Üben.

Je besser die Einzelbewertung der sieben Kriterien ausfällt, desto größer ist die Wahrscheinlichkeit Unterrichtsqualität höher einzuschätzen. Die Bewertung des Kriteriums „Sinnstiftende Kommunikation“ übte den größten Einfluss aus - ein Anstieg um eine Rangstufe ist mit einer 3fach erhöhten Chance verbunden Unterrichtsqualität eine Stufe höher zu beurteilen. Für die Kriterien „Lernförderliches Klima“, „Methodenvielfalt“ sowie „Vorbereitete Unterrichtsumgebung“ konnte kein signifikanter Einfluss auf die studentische Globalbewertung von Unterrichtsqualität festgestellt werden.

Tabelle 8: Koeffizienten und Gütekriterien der ordinalen Regression zur Bestimmung des Einflusses der zehn Kriterien auf die Globalbewertung von Unterrichtsqualität (n=256)

Kriterien	Koeffizient	SE	Signifikanz	Odds Ratio	[95% KI]
Sinnstiftende Kommunikation	1,08	0,21	<0,001**	2,96	[1,93; 4,52]
Inhaltliche Klarheit	0,74	0,22	0,001**	2,1	[1,35; 3,26]
Klare Strukturierung	0,54	0,19	0,005**	1,72	[1,15; 2,56]
Individuelles Fördern	0,54	0,16	0,001**	1,72	[1,24; 2,39]
Hoher Anteil echter Lernzeit	0,53	0,18	0,004**	1,7	[1,19; 2,44]
Transparente Leistungserwartungen	0,38	0,19	0,042*	1,47	[1,00; 2,14]
Effektives Üben	0,38	0,17	0,026*	1,46	[1,04; 2,06]
Lernförderliches Klima	0,24	0,22	0,259	1,28	[0,82; 1,99]
Methodenvielfalt	0,16	0,16	0,319	1,17	[0,85; 1,63]
Vorbereitete Unterrichtsumgebung	-0,18	0,19	0,349	0,84	[0,57; 1,23]
Likelihood ratio (10) = 316,59 (p<0,001**)					
McFadden Pseudo R^2 = 0,48					
Cox & Snell Pseudo R^2 = 0,71					
Nagelkerke Pseudo R^2 = 0,77					
% korrekte Klassifikationen = 72,27% > PZW = 32%					

*p<0,05, ** p<0,01, SE: Standardfehler, KI: Konfidenzintervall

Die Güte des Regressionsmodells kann anhand der in Abschnitt 3.4.4. genannten Gütekriterien als sehr hoch eingeschätzt werden. Zwischen den zehn Kriterien bestand keine relevante Multikollinearität, die im Modell hätte berücksichtigt werden müssen. Die VIF-Werte waren kleiner 10 und die Konditionsindices lagen unter 30. Der Parallelitätstest für Linien war nicht signifikant (p=0,676), so dass die Voraussetzung zur Durchführung einer ordinalen Regressionsanalyse erfüllt waren.

4.4. Einfluss des Unterrichtsformats

Zur Untersuchung des Einflusses des Unterrichtsformats (UaK oder Seminar) wurde ein ordinales Regressionsmodell mit der Gesamtstichprobe berechnet. Dabei wurde eine binäre Dummy Variable „Unterrichtsformat“ (UF) für die Gruppenzugehörigkeit (0=Referenzgruppe=UaK; 1=Vergleichsgruppe=Seminar) sowie Interaktionsterme als Produkte der Gruppenzugehörigkeit und der einzelnen zehn Kriterien eingefügt. Der Parallelitätstest für Linien war nicht signifikant (p=0,128), so dass die Voraussetzung zur Durchführung einer ordinalen Regressionsanalyse erfüllt waren.

Zusätzlich zum ordinalen Regressionsmodell (Tabelle 9) wurde ein Heterogeneous Choice Modell (Tabelle 10) berechnet, das für die Gruppierungsvariable „Unterrichtsformat“ Heteroskedastizität zulässt.

Beide Regressionsmodelle zeigten im Vergleich der beiden Unterrichtsformate einen signifikanten Unterschied des Einflusses der beiden Kriterien „Inhaltliche Klarheit“ und „Individuelles Fördern“ auf die Globalbewertung von Unterrichtsqualität. Die Interaktionsterme für „Inhaltliche Klarheit“ und „Unterrichtsformat“ sowie „Individuelles Fördern“ und „Unterrichtsform“ waren statistisch signifikant. Der Einfluss des Kriteriums „Inhaltliche Klarheit“ auf die Globalbeurteilung von Unterrichtsqualität war im Unterrichtsformat „Seminar“ signifikant niedriger als im Unterrichtsformat „Unterricht am Krankenbett“. Dahingegen war der Einfluss des Kriteriums „Individuelles Fördern“ im Unterrichtsformat „Seminar“ signifikant höher.

Im Vergleich zwischen dem Heterogeneous Choice Modell und dem ordinalen Regressionsmodell konnte anhand des niedrigen BIC-Wertes dem Heterogeneous Choice Modell Vorrang gegeben werden (BIC 464,45 vs. BIC 465,07). Das Heterogeneous Choice Modell bestätigte die vermutete Heteroskedastizität zwischen den beiden Gruppen der Unterrichtsformate ($\ln \delta$ UF = -0,45, p=0,015).

Tabelle 9: Koeffizienten und Gütekriterien des Heterogeneous Choice Modells zur Bestimmung des Einflusses der zehn Kriterien auf die Globalbewertung von Unterrichtsqualität im Vergleich der Unterrichtsformate „Unterricht am Krankenbett“ und „Seminar“ (n=256, n UaK = 75, n Seminar = 181)

Kriterien	Koeffizient	SE	Signifikanz	Odds Ratio	[95% KI]
Klare Strukturierung	0,55	0,39	0,164	1,73	[0,8; 3,74]
Hoher Anteil echter Lernzeit	0,29	0,29	0,313	1,34	[0,76; 2,35]
Lernförderliches Klima	0,29	0,45	0,513	1,34	[0,56; 3,21]
Inhaltliche Klarheit	1,39	0,45	0,002**	4,01	[1,64; 9,77]
Sinnstiftende Kommunikation	0,26	0,35	0,456	1,3	[0,65; 2,59]
Methodenvielfalt	0,3	0,31	0,322	1,36	[0,74; 2,48]
Individuelles Fördern	-0,16	0,33	0,631	0,85	[0,45; 1,63]
Effektives Üben	0,66	0,36	0,066	1,94	[0,96; 3,93]
Transparente Leistungserwartungen	-0,1	0,39	0,799	0,9	[0,42; 1,96]
Vorbereitete Unterrichtsumgebung	-0,05	0,31	0,874	0,95	[0,52; 1,75]
Dummy Unterrichtsformat	-0,87	2,02	0,669	0,42	[0,01; 22,3]
Klare Strukturierung x UF	-0,12	0,43	0,784	0,89	[0,39; 2,05]
Hoher Anteil echter Lernzeit x UF	0,13	0,33	0,686	1,14	[0,6; 2,16]
Lernförderliches Klima x UF	-0,09	0,49	0,855	0,91	[0,35; 2,38]
Inhaltliche Klarheit x UF	-1,05	0,47	0,026*	0,35	[0,14; 0,88]
Sinnstiftende Kommunikation x UF	0,62	0,4	0,123	1,85	[0,85; 4,05]
Methodenvielfalt x UF	-0,3	0,34	0,368	0,74	[0,38; 1,43]
Individuelles Fördern x UF	0,76	0,38	0,042*	2,14	[1,03; 4,48]
Effektives Üben x UF	-0,43	0,38	0,258	0,65	[0,31; 1,37]
Transparente Leistungserwartungen x UF	0,56	0,43	0,194	1,75	[0,75; 4,08]
Vorbereitete Unterrichtsumgebung x UF	0,02	0,37	0,962	1,02	[0,5; 2,08]
ln δ UF	-0,45	0,19	0,015*	0,64	[0,44; 0,92]
Likelihood ratio (22) = 343,99 (p<0,001)					
McFadden Pseudo R^2 = 0,52					
Cox & Snell Pseudo R^2 = 0,74					
Nagelkerke Pseudo R^2 = 0,8					
AIC = 372,27					
BIC = 464,45					

*p<0,05, ** p<0,01, UF: Unterrichtsformat, SE: Standardfehler, KI: Konfidenzintervall

Tabelle 10: Koeffizienten und Gütekriterien des ordinalen Regressionsmodells zur Bestimmung des Einflusses der zehn Kriterien auf die Globalbewertung von Unterrichtsqualität im Vergleich der Unterrichtsformate „Unterricht am Krankenbett“ und „Seminar“ (n=256, n UaK = 75, n Seminar = 181)

Kriterien	Koeffizient	SE	Signifikanz	Odds Ratio	[95% KI]
Klare Strukturierung	0,67	0,44	0,129	1,95	[0,82; 4,65]
Hoher Anteil echter Lernzeit	0,4	0,32	0,202	1,5	[0,81; 2,78]
Lernförderliches Klima	0,36	0,49	0,468	1,43	[0,54; 3,76]
Inhaltliche Klarheit	1,75	0,47	<0,001**	5,75	[2,31; 14,3]
Sinnstiftende Kommunikation	0,41	0,39	0,294	1,5	[0,7; 3,2]
Methodenvielfalt	0,39	0,34	0,243	1,48	[0,77; 2,86]
Individuelles Fördern	-0,2	0,37	0,594	0,82	[0,39; 1,7]
Effektives Üben	0,9	0,4	0,025*	2,46	[1,12; 5,42]
Transparente Leistungserwartungen	-0,14	0,43	0,746	0,87	[0,37; 2,04]
Vorbereitete Unterrichtsumgebung	-0,07	0,34	0,824	0,93	[0,47; 1,81]
Dummy Unterrichtsformat	-2,36	2,32	0,31	0,09	[0; 8,97]
Klare Strukturierung x UF	-0,06	0,5	0,905	0,94	[0,35; 2,53]
Hoher Anteil echter Lernzeit x UF	0,18	0,4	0,657	1,19	[0,55; 2,59]
Lernförderliches Klima x UF	-0,07	0,57	0,909	0,94	[0,31; 2,86]
Inhaltliche Klarheit x UF	-1,28	0,53	0,015*	0,28	[0,1; 0,78]
Sinnstiftende Kommunikation x UF	0,83	0,47	0,077	2,3	[0,91; 5,78]
Methodenvielfalt x UF	-0,39	0,4	0,323	0,68	[0,31; 1,47]
Individuelles Fördern x UF	1,04	0,43	0,0017*	2,82	[1,2; 6,58]
Effektives Üben x UF	-0,57	0,45	0,201	0,56	[0,23; 1,36]
Transparente Leistungserwartungen x UF	0,76	0,49	0,127	2,13	[0,81; 5,6]
Vorbereitete Unterrichtsumgebung x UF	0,05	0,44	0,918	1,05	[0,44; 2,5]
Likelihood ratio (21) = 337,82 (p<0,001**)					
McFadden Pseudo R^2 = 0,51					
Cox & Snell Pseudo R^2 = 0,73					
Nagelkerke Pseudo R^2 = 0,79					
AIC = 376,44					
BIC = 465,07					

*p<0,05, ** p<0,01, UF: Unterrichtsformat, SE: Standardfehler, KI: Konfidenzintervall

4.5. Einfluss des Zeitpunktes im Studienverlauf

Zur Untersuchung eines möglichen Einflusses des Zeitpunktes im Studienverlauf (7. versus 9. Fachsemester) wurde ein ordinales Regressionsmodell mit der Gesamtstichprobe berechnet. Dabei wurde eine binäre Dummy Variable „Semester“ (Sem) für die Gruppenzugehörigkeit (0 = Referenzgruppe = 7. Semester; 1 = Vergleichsgruppe = 9. Semester) sowie Interaktionsterme als Produkte der Gruppenzugehörigkeit und der einzelnen zehn Kriterien eingefügt.

Der Parallelitätstest für Linien war nicht signifikant (p=0,322), so dass die Voraussetzung zur Durchführung einer ordinalen Regressionsanalyse erfüllt waren.

Zusätzlich zum ordinalen Regressionsmodell (Tabelle 11) wurde ein Heterogeneous Choice Modell (Tabelle 12), das für die Gruppierungsvariable „Semester“ Heteroskedastizität zulässt, berechnet.

Beide Regressionsmodelle zeigten im Vergleich der beiden Zeitpunkte im Studienverlauf keinen signifikanten Unterschied des Einflusses der zehn Kriterien. Keiner der Interaktionsterme der einzelnen Kriterien und der Dummy Variabel „Semester“ waren statistisch signifikant.

Im Vergleich zwischen dem Heterogeneous Choice Modell und dem ordinalen Regressionsmodell konnte anhand des niedrigeren BIC-Wertes dem ordinalen Regressionsmodell Vorrang gegeben werden (BIC 471,11 vs. BIC 475,1). Das Heterogeneous Choice Modell zeigte keine Heteroskedastizität zwischen den beiden Gruppen der untersuchten Semester ($\ln \delta$ Sem = 0,21, p=0,21).

Tabelle 11: Koeffizienten und Gütekriterien des Heterogeneous Choice Modells zur Bestimmung des Einflusses der zehn Kriterien auf die Globalbewertung von Unterrichtsqualität im Vergleich der zwei Zeitpunkte im Studienverlauf „7. Semester“ und „9. Semester“ (n=256, n 7. Semester = 121, n 9. Semester = 135)

Kriterien	Koeffizient	SE	Signifikanz	Odds Ratio	[95% KI]
Klare Strukturierung	0,81	0,38	0,032*	2,24	[1,1; 4,69]
Hoher Anteil echter Lernzeit	0,43	0,31	0,168	1,54	[0,83; 2,83]
Lernförderliches Klima	0,69	0,37	0,062	2	[0,97; 4,14]
Inhaltliche Klarheit	1,17	0,4	0,004**	3,22	[1,46; 7,08]
Sinnstiftende Kommunikation	1,12	0,35	0,001**	3,01	[1,56; 6,09]
Methodenvielfalt	0,44	0,26	0,086	1,56	[0,94; 2,59]
Individuelles Fördern	0,84	0,31	0,007**	2,32	[1,26; 4,26]
Effektives Üben	0,15	0,26	0,564	1,17	[0,69; 1,96]
Transparente Leistungserwartungen	0,08	0,33	0,797	1,09	[0,57; 2,07]
Vorbereitete Unterrichtsumgebung	-0,16	0,33	0,623	0,85	[0,45; 1,62]
Dummy Semester	3,6	2,24	0,108	36,55	[0,45; 2954]
Klare Strukturierung x Sem	-0,19	0,49	0,703	0,83	[0,31; 2,15]
Hoher Anteil echter Lernzeit x Sem	0,21	0,42	0,624	1,23	[0,54; 2,81]
Lernförderliches Klima x Sem	-0,82	0,52	0,12	0,44	[0,16; 1,24]
Inhaltliche Klarheit x Sem	-0,19	0,52	0,707	0,82	[0,3; 2,27]
Sinnstiftende Kommunikation x Sem	0,14	0,48	0,771	1,15	[0,45; 2,91]
Methodenvielfalt x Sem	-0,68	0,41	0,102	0,51	[0,23; 1,14]
Individuelles Fördern x Sem	-0,47	0,4	0,243	0,62	[0,28; 1,38]
Effektives Üben x Sem	0,74	0,42	0,083	2,01	[0,91; 4,8]
Transparente Leistungserwartungen x Sem	0,53	0,45	0,239	1,7	[0,7; 4,12]
Vorbereitete Unterrichtsumgebung x Sem	-0,18	0,46	0,705	0,84	[0,34; 2,08]
ln δ UF	0,21	0,17	0,211	0,21	[-0,12; 0,53]
Likelihood ratio (21) = 333,34 (p<0,001)					
McFadden Pseudo R^2 = 0,5					
Cox & Snell Pseudo R^2 = 0,73					
Nagelkerke Pseudo R^2 = 0,79					
AIC = 382,93					
BIC = 475,1					

*p<0,05, ** p<0,01, Sem: Semester, SE: Standardfehler, KI: Konfidenzintervall

Tabelle 12: Koeffizienten und Gütekriterien des ordinalen Regressionsmodells zur Bestimmung des Einflusses der zehn Kriterien auf die Globalbewertung von Unterrichtsqualität im Vergleich der zwei Zeitpunkte im Studienverlauf „7. Semester“ und „9. Semester“ (n=256, n 7. Semester = 121, n 9. Semester = 135)

Kriterien	Koeffizient	SE	Signifikanz	Odds Ratio	[95% KI]
Klare Strukturierung	0,73	0,35	0,04*	2,07	[1,04; 4,15]
Hoher Anteil echter Lernzeit	0,37	0,29	0,209	1,44	[0,81; 2,56]
Lernförderliches Klima	0,62	0,34	0,073	1,85	[0,94; 3,64]
Inhaltliche Klarheit	1,03	0,37	0,005**	2,8	[1,37; 5,77]
Sinnstiftende Kommunikation	1,01	0,32	0,001**	2,74	[1,47; 5,1]
Methodenvielfalt	0,39	0,24	0,106	1,48	[0,92; 2,37]
Individuelles Fördern	0,75	0,29	0,009**	2,13	[1,21; 3,73]
Effektives Üben	0,14	0,25	0,558	1,16	[0,71; 1,89]
Transparente Leistungserwartungen	0,06	0,31	0,84	1,06	[0,58; 1,96]
Vorbereitete Unterrichtsumgebung	-0,14	0,32	0,651	0,87	[0,47; 1,61]
Dummy Semester	3,51	1,99	0,078	33,51	[0,67; 1666]
Klare Strukturierung x Sem	-0,19	0,44	0,666	0,83	[0,35; 1,96]
Hoher Anteil echter Lernzeit x Sem	0,19	0,38	0,61	1,21	[0,58; 2,56]
Lernförderliches Klima x Sem	-0,73	0,46	0,114	0,48	[0,19; 1,19]
Inhaltliche Klarheit x Sem	-0,18	0,46	0,696	0,83	[0,33; 2,08]
Sinnstiftende Kommunikation x Sem	0,09	0,42	0,823	1,1	[0,48; 2,53]
Methodenvielfalt x Sem	-0,59	0,36	0,1	0,55	[0,27; 1,12]
Individuelles Fördern x Sem	-0,44	0,37	0,23	0,65	[0,32; 1,32]
Effektives Üben x Sem	0,63	0,37	0,085	1,88	[0,92; 3,88]
Transparente Leistungserwartungen x Sem	0,47	0,4	0,241	1,6	[0,73; 3,53]
Vorbereitete Unterrichtsumgebung x Sem	-0,16	0,42	0,702	0,85	[0,37; 1,93]
Likelihood ratio (21) = 331,78 (p<0,001**)					
McFadden Pseudo R^2 = 0,5					
Cox & Snell Pseudo R^2 = 0,73					
Nagelkerke Pseudo R^2 = 0,79					
AIC = 382,49					
BIC = 471,11					

*p<0,05, ** p<0,01, Sem: Semester, SE: Standardfehler, KI: Konfidenzintervall

5. DISKUSSION

5.1. Relevanz der zehn Kriterien des BTQ-10 für die Globalbewertung von Unterrichtsqualität

In der vorliegenden Arbeit wurde ein neues Beobachtungsinstrument zur Bewertung von Unterrichtsqualität aus studentischer Perspektive in der medizinischen Ausbildung entwickelt und erfolgreich validiert - der Berlin Teaching Quality Questionnaire 10 (BTQ-10). Im Vergleich zu bisher publizierten Instrumenten verfügen die im BTQ-10 verwendeten Kriterien über eine breite empirische Basis, die sich auf umfangreiche Arbeiten der Allgemeinen Pädagogik stützt. Viele der bisher verwendeten Instrumente untersuchen darüber hinaus den Bereich der Facharztausbildung. Wie auch Silber et al. darlegen, unterscheiden sich jedoch Studierende und Assistenzärzte darin, was sie für essentielle Dimensionen von guter Lehre erachten bzw. benötigen [36]. So steht für Assistenzärzte vielmehr die klinische Kompetenz des Lehrers im Vordergrund [36,41,144]. Des Weiteren war die Übertragbarkeit der bereits verfügbaren Instrumente auf das deutsche Medizinstudium nur eingeschränkt möglich, da ein Großteil der Messinstrumente im angelsächsischen Sprachraum entwickelt und validiert wurde [24]. Mit dem BTQ-10 steht erstmalig ein an einer deutschen medizinischen Fakultät entwickelter und validierter Fragebogen zur Verfügung.

Um die Validität des BTQ-10 zu überprüfen, wurde das theoretische Konstrukt Unterrichtsqualität im Rahmen eines formativen Messmodells operationalisiert. In keiner Arbeit zu bisher publizierten Instrumenten zur Bewertung von Unterrichtsqualität findet sich eine Unterscheidung zwischen den beiden Operationalisierungsansätzen. Wie jedoch bereits in Abschnitt 2.5.3. dargelegt, können Fehler in der Wahl der Indikator-Konstrukt-Beziehung zu fehlerhaften Untersuchungsergebnissen und falschen Schlüssen führen. Zumeist wurde von anderen Autoren ein reflektiver Ansatz verfolgt, ohne dass dieser jedoch begründet wurde. Somit finden sich einige Arbeiten, in denen eine fehlerhafte Operationalisierung vorliegt [26, 38]. Viele Konstrukte können je nach Kontext sowohl formativ als auch reflektiv operationalisiert werden; die Entscheidung zwischen reflektiven und formativen Messmodellen sollte dabei jedoch immer genau hinterfragt werden [90, 91]. Dazu können die im Abschnitt 2.5.3. angeführten Entscheidungskriterien benutzt werden.

Zur Untersuchung des kausalen Zusammenhanges zwischen den zehn Kriterien und der Globalbewertung von Unterrichtsqualität und zum Nachweis der Validität des BTQ-10 wurde in der vorliegenden Arbeit eine ordinale Regressionsanalyse durchgeführt. Einige Autoren haben

einzig Korrelationsanalysen verwendet, um die Validität ihrer Instrumente zu bestätigen [51,53,45]. Die Korrelation gibt jedoch noch keinen Anhalt für einen kausalen Zusammenhang und ist zur Bestätigung, dass das Konstrukt durch die gewählten Indikatoren valide abgebildet wird, unzureichend [76]. In der durchgeführten Regressionsanalyse zeigte sich, dass die drei Kriterien „Lernförderliches Klima“, „Methodenvielfalt“ und „Vorbereitete Unterrichtsumgebung“ keine signifikanten Einflussfaktoren für die Globalbewertung von Unterrichtsqualität darstellten.

Da die zehn Kriterien aus der Allgemeinen Pädagogik abgeleitet wurden, kann dies ein Hinweis darauf sein, dass die genannten drei Kriterien im Bereich der Hochschulpädagogik, insbesondere im Rahmen der medizinischen Ausbildung, einen weniger relevanten Einfluss aufweisen. Auf der anderen Seite ist dabei jedoch zu beachten, dass die Ergebnisse auf studentische Bewertungen zurückgehen und somit diese Kriterien möglicherweise für die Bewertung von Unterrichtsqualität durch Studierende keine Relevanz besitzen. Um den Einfluss der drei Kriterien auf die Globalbewertung von Unterrichtsqualität genauer zu untersuchen, könnte der BTQ-10 in weiterführenden Untersuchungen von Dozierenden und externen Beobachtern angewandt werden. Stellen sich die Kriterien in diesem Rahmen als signifikante Einflussfaktoren heraus, würde das die Hypothese untermauern, dass die genannten drei Kriterien in der Globalbewertung von Unterrichtsqualität durch Studierende weniger relevant sind.

Irby et al. [65] und MacDonald et al. [145] konnten in diesem Zusammenhang zeigen, dass die Unterrichtsumgebung für Studierende wenig Einfluss auf die Bewertung von Unterrichtsqualität hat. Dazu korrelierten Irby et al. mehrere Faktoren der Unterrichtsumgebung mit der Gesamtbewertung der Lehre. Die Faktoren umfassten “case mix of clinic“, „work load“, „structured time for teaching“, „space for teaching” und „organization of clinic”. Die Punkte „space for teaching“ und „organization of clinic” korrespondieren dabei mit dem Kriterium “Vorbereitete Unterrichtsumgebung” des BTQ-10. Keiner der genannten Faktoren korrelierte mit der Gesamtbewertung. Auch in der multiplen Regressionsanalyse stellte sich keiner dieser Punkte als signifikanter Einflussfaktor heraus. Dies unterstützt die Vermutung, dass das Kriterium „Vorbereitete Unterrichtsumgebung“ aus der Perspektive der Studierenden keine relevante Rolle spielt.

Es sollte jedoch auch die Formulierung der typisch beobachtbaren Beispiele der drei Kriterien im BTQ-10 kritisch überprüft werden. Möglicherweise wurden diese nicht ausreichend verständlich

formuliert und gaben den Inhalt der Kriterien (siehe Abschnitt 2.2.) nicht suffizient wieder. Dies könnte insbesondere bei dem Kriterium „Lernförderliches Klima“ relevant sein.

Betrachtet man in diesem Zusammenhang den SFDP26-German (Tabelle 14), so lassen sich mehrere Punkte im Vergleich zum BTQ-10 feststellen. Iblher et al. haben in ihrer Untersuchung die sieben Skalen des SFDP26-German mit der Gesamtlehrereffektivität korreliert. Am stärksten war der Zusammenhang mit den Skalen „Zielkommunikation“ (r=0,61), „Fördern von Verstehen und Behalten“ (r=0,58) sowie „Etablieren des Lernklimas“ (r=0,51). Für die vier übrigen Skalen lagen die Korrelationskoeffizienten zwischen 0,4 und 0,5 [50].

Tabelle 13: Darstellung des SFDP26-German [50] und Zuordnung zu den Kriterien des BTQ-10

SFDP26-German		Kriterien BTQ-10
Skalen	Items	
Etablieren des Lernklimas	Er/sie hörte den Lernenden zu.	
	Er/sie bestärkte die Lernenden, sich aktiv an der Diskussion zu beteiligen.	Individuelles Fördern
	Er/sie verhielt sich respektvoll den Lernenden gegenüber.	Lernförderliches Klima
	Er/Sie ermutigte die Lernenden, Probleme anzusprechen	Lernförderliches Klima
Leitung einer Lehreinheit	Er/sie achtete auf den zeitlichen Rahmen.	Hoher Anteil echter Lernzeit
	Er/sie vermied Abschweifungen vom Kursthema.	
	Er/sie verhinderte äußere Störungen.	
Zielkommunikation	Er/sie stellte die Lernziele des Kurses kurz und prägnant dar.	Transparente Leistungserwartungen
	Er/sie machte den Lernenden die Relevanz der Lernziele deutlich.	Sinnstiftende Kommunikation
	Er/sie setzte inhaltliche Schwerpunkte (Prioritäten im Kurs).	Inhaltliche Klarheit
	Er/sie wiederholte die Kurslernziele regelmäßig.	Klare Strukturierung
Fördern von Verstehen und Behalten	Er/sie präsentierte den Lernstoff gut strukturiert.	Klare Strukturierung
	Er/sie machte die Zusammenhänge des Lernstoffs deutlich.	
	Er/sie setzte visuelle Hilfen zur Veranschaulichung ein (Tafel, Modelle).	Methodenvielfalt

Evaluation	Er/sie überprüfte medizinisches Vorwissen der Lernenden. Er/sie überprüfte Fähigkeiten der Lernenden, medizinische Zusammenhänge herzustellen. Er/sie überprüfte Fähigkeiten der Lernenden, medizinisches Wissen patientenspezifisch anzuwenden. Er/sie überprüfte praktische Fertigkeiten der Lernenden in der fallbezogenen Patientenversorgung.	Inhaltliche Klarheit
Feedback	Er/sie gab den Lernenden negatives (korrigierendes) Feedback. Er/sie erklärte den Lernenden, was richtig und was falsch war. Er/sie machte den Lernenden Vorschläge, wie sie es besser machen könnten. Er/sie gab regelmäßig Feedback an die Lernenden.	Transparente Leistungserwartungen
Fördern von selbst-bestimmtem Lernen	Er/sie bestärkte die Lernenden ausdrücklich, das Themengebiet weiter zu vertiefen. Er/sie motivierte die Lernenden, selbstständig zu lernen. Er/sie bestärkte die Lernenden, außerhalb des Kurses nachzulesen.	
Im Ganzen war die Lehrleistung des Dozenten ...		

Das Ergebnis für die Skala „Etablieren des Lernklimas“ steht im Gegensatz zu den Ergebnissen für den BTQ-10. Die beiden Items „Er/sie verhielt sich respektvoll den Lernenden gegenüber.“ sowie „Er/sie ermutigte die Lernenden, Probleme anzusprechen.“ der Skala „Etablieren des Lernklimas“ finden sich inhaltlich in dem Kriterium „Lernförderliches Klima“ des BTQ-10 wieder (siehe Abschnitt 2.2.). Es war anzunehmen, dass sich das Kriterium „Lernförderliches Klima“ ebenso als relevanter Einflussfaktor nachweisen lässt. Das konnte jedoch nicht bestätigt werden. Dies könnte ein Hinweis darauf sein, dass dieses Kriterium im BTQ-10 nicht spezifisch genug beschrieben wurde. Es wies auch den stärksten Deckeneffekt der zehn Kriterien auf (siehe Abbildung 8), was diese These unterstützt.

Für das Kriterium „Vorbereitete Unterrichtsumgebung“ des BTQ-10 lassen sich keine im SFDP26-German vergleichbare Items finden. Das Item „Er/sie setzte visuelle Hilfen zur

Veranschaulichung ein (Tafel, Modelle, etc.)." lässt sich inhaltlich grob dem Kriterium „Methodenvielfalt" zuordnen. Dazu ist jedoch anzumerken, dass dieses Item zusammen mit dem Item „Er/sie präsentiert den Lernstoff gut strukturiert." und „Er/sie macht die Zusammenhänge des Lernstoffs deutlich." in die Skala „Fördern von Verstehen und Behalten" eingeht. In der Untersuchung von Ilbher wurden die Skalengesamtwerte als Mittelwerte der Items berechnet und anschließend mit der Gesamtlehrereffektivität korreliert. Somit könnte der möglichweise geringere Zusammenhang des einzelnen Items „Er/sie setzte visuelle Hilfen zur Veranschaulichung ein (Tafel, Modelle, etc.)." mit der Gesamtlehrereffektivität in der Bildung des Mittelwerts mit den beiden anderen Items der Skala verschleiert worden sein.

Zusammenfassend lässt sich somit festhalten, dass die beiden Kriterien „Vorbereitete Unterrichtsumgebung" und „Methodenvielfalt" für die Anwendung des BTQ-10 zur Bewertung von Unterrichtsqualität aus studentischer Perspektive geringere Relevanz besitzen und für die weitere Anwendung des BTQ-10 gestrichen werden könnten. Der Einfluss des Kriteriums „Lernförderliches Klima" sollte erneut mit einer präziseren Formulierung der Aussagen zu dem Kriterium im BTQ-10 untersucht werden.

Für Studierende scheint es am wichtigsten zu sein, dass ihnen der Nutzen des Kurses für ihre weitere Ausbildung klar war. Dieser Punkt war im Kriterium „Sinnstiftende Kommunikation" enthalten und stellte sich als größter Einflussfaktor für die Globalbewertung der Unterrichtsqualität heraus. Es folgten die Kriterien „Inhaltliche Klarheit", „Klare Strukturierung" und „Individuelles Fördern". Diese vier Kriterien lassen sich in den Skalen des SFDP26-German wiederfinden, die den stärksten Zusammenhang zur Gesamtlehrereffektivität aufwiesen (vergleichbar Tabelle 13). Des Weiteren lassen sich die Items der im SFDP26-German führenden Skala „Zielkommunikation" in inhaltlichen Facetten von vier verschiedenen signifikanten Kriterien des BTQ-10 wiederfinden. Dies unterstreicht die Relevanz und Anwendbarkeit der empirisch basierten, aus der Allgemeinen Pädagogik entlehnten Kriterien des BTQ-10.

5.2. Einfluss des Unterrichtsformats

In der vorliegenden Arbeit zeigten sich signifikante Unterschiede zwischen den beiden untersuchten Unterrichtsformaten im Einfluss der zehn Kriterien auf die Globalbewertung von Unterrichtsqualität. Das traf auf die Kriterien „Inhaltliche Klarheit" und „Individuelles Fördern" zu.

Im Unterrichtsformat „Seminar", das eher durch einen frontalen Unterrichtsstil charakterisiert ist und in dem in den untersuchten Lehrveranstaltungen eine doppelt bis dreifach so hohe mittlere

Gruppenstärke vorlag (siehe Tabelle 6), wurde dem Kriterium „Individuelles Fördern“ ein höherer Stellenwert beigemessen als im Unterrichtsformat „Unterricht am Krankenbett“. Demgegenüber konnte für das Unterrichtsformat „UaK“ gezeigt werden, dass das Kriterium „Inhaltliche Klarheit“ einen signifikant stärkeren Einfluss aufwies.

Dies zeigt, dass Studierende unterschiedliche edukative Bedürfnisse in Abhängigkeit vom Unterrichtsformat haben. Somit sollten Dozierende in Seminaren darauf achten, alle Studierenden in den Unterricht mit einzubeziehen und den individuellen Lernfortschritt kontinuierlich zu monitoren und gegebenenfalls entsprechende Hilfestellungen anzubieten. Das Kriterium „Inhaltliche Klarheit“ ist unter anderem durch die Aussagen „Die gestellten Aufgaben waren inhaltlich gut verständlich.“ und „Der Dozent hat das bereits vorhandene Wissen der Studierenden in Erfahrung gebracht und an dieses inhaltlich angeknüpft.“ im BTQ-10 beschrieben. Demzufolge sollten Dozierende insbesondere im Unterricht am Krankenbett, in dem eine Patientenuntersuchung und nachfolgende Patientenvorstellung im Vordergrund steht, überprüfen, ob die genannten Punkte erfüllt sind. Studierende nehmen die Patientenuntersuchung häufig ohne den Dozierenden allein oder in Kleingruppen vor. Somit kann eine ungenaue Aufgabenstellung, aber auch das fehlende Rüstzeug im Sinne von Wissen und Fertigkeiten zu einer für die Studierenden ungenügsamen Unterrichtserfahrung werden.

5.3. Einfluss des Zeitpunktes im Studienverlauf

Es war kein signifikanter Unterschied in den Einflussstärken der zehn Kriterien auf die Globalbewertung von Unterrichtsqualität zwischen den beiden Zeitpunkten im Studienverlauf nachweisbar. Fluit et al. [8] und Beckman et al. [22] weisen zwar darauf hin, dass sich Lernenden auf unterschiedlichen Stufen in dem unterscheiden, was sie an Lehrern schätzen. Dieses bezieht sich jedoch zumeist auf den Unterschied zwischen Studierenden und Assistenzärzten [48, 146] oder zwischen Assistenzärzten zu unterschiedlichen Zeitpunkten in ihrer Facharztausbildung [37, 89]. Wie bereits dargestellt, stellt für Assistenzärzte im Vergleich zu Studierenden die klinische Kompetenz des Lehrers ein entscheidendes Bewertungsmerkmal dar [36, 41, 48, 144, 146]. Je weiter Assistenzärzte in ihrer Facharztausbildung fortschreiten, desto mehr betrachten sie ihre klinischen Lehrer als Kollegen und messen der Eigenverantwortlichkeit für das persönliche Lernen eine größere Bedeutung bei [37, 89].

5.4. Methodendiskussion – und kritik

Für die Messung der formativen Indikatoren sowie der Gesamtbeurteilung der Unterrichtsqualität innerhalb des BTQ-10 wurden fünfstufige Ratingskalen verwendet.

Antworttendenzen, zum Beispiel Neigung zu extremen Antworten oder die Tendenz zu mittleren Urteilen [77, 78], die als Nachteil von Ratingskalen auftreten können, waren in den untersuchten 256 Fragebögen nicht festzustellen. Jedoch zeigte sich für alle zehn Kriterien sowie die Globalbewertung der Unterrichtsqualität ein Deckeneffekt (siehe Abbildungen 7-17), der auf eine nicht ausreichende Differenzierbarkeit im oberen Bereich schließen lässt. Dieses wurde auch von anderen Autoren beobachtet, so dass Zuberi et al. [40] eine siebenstufige Ratingskala mit nach links verschobener Mitte verwendeten.

Die verbale Verankerung der Ratingskala wurde bewusst von +2 nach -2 gewählt, um eine versehentliche Umkehr beim Ankreuzen durch eine Verwechslung mit dem Schulnotennotensystem zu vermeiden [50].

Die Beurteilung der Unterrichtqualität mittels des BTQ-10 hätte durch die Kooperationsbereitschaft der Studierenden an der Beurteilung mitzuwirken beeinflusst werden können. Um Verständnis zu erzeugen und die Motivation zu erhöhen, wurde vor Beginn der Lehrveranstaltung eine kurze Einführung zum Ziel der Untersuchung gegeben und die Studierenden wurden gebeten, den BTQ-10 gewissenhaft auszufüllen. Die Rücklaufquote von 92,7% zeigt eine hohe Beteiligung der Studierenden. Insbesondere in den Lehrveranstaltungen der Unterrichtsform „Unterricht am Krankenbett“ mit kleinen Studierendengruppen konnte eine Rücklaufquote von 100% erreicht werden. In den Freitexten des BTQ-10 fanden sich positive Rückmeldungen zum Ziel der Untersuchung, die auf eine hohe innere Bereitschaft der Studierenden, an der Untersuchung qualitativ mitzuwirken, schließen lässt.

Diese Studie hat verschiedene Limitationen.

Zwischen den beiden Zeitpunkten im Studienverlauf zeigte sich kein signifikanter Unterschied. Möglicherweise ist jedoch der Unterschied zwischen den gewählten Semestern zu gering. In einer folgenden Untersuchung könnte man einen größeren Unterschied zwischen den beiden Zeitpunkten im Studienverlauf wählen, um gegebenenfalls einen Unterschied in den Einflussstärken der zehn Kriterien zu detektieren. Es wurden keine personenspezifischen Daten der Studierenden erhoben, so dass keine Angaben zum Alter der Studierenden in den beiden Semestern gemacht werden konnten.

Die Reliabilität des BTQ-10 wurde nicht untersucht. Reliabilität ist neben Validität ein wichtiges Gütekriterium [74-77]. Wie in Abschnitt 2.5.2.1. beschrieben, ist es für formative Messmodelle möglich die Retest-Reliabiltät (Wiederholungsreliabilität) zu bestimmen. Dies wäre durch Videoaufzeichnungen von Unterrichtseinheiten denkbar, könnte jedoch beim Beobachter zu Erinnerungs- und Lerneffekten führen, die die Ergebnisse verfälschen würden [74].

Der BTQ-10 kann für unterschiedliche Unterrichtsformate eingesetzt werden. Jedoch sollte die Übertragbarkeit auf andere Länder außerhalb des deutschen Sprachraums überprüft werden.

5.5. Schlussfolgerung

Die aus der allgemeinen Pädagogik entlehnten Kriterien verfügen über ein breites empirisches Fundament und weisen damit im Gegensatz zu bisher publizierten Instrumenten zur Erfassung von Unterrichtsqualität einen deutlichen Vorteil auf. Es konnte in der vorliegenden Untersuchung gezeigt werden, dass sie in der Lage sind, Unterrichtsqualität in der medizinischen Ausbildung aus studentischer Perspektive valide zu erfassen. Im Rahmen der Arbeit war es möglich, einen Einblick in die Bewertungsmaßstäbe für guten Unterricht aus der Sicht von Studierenden zu erlangen und zu untersuchen, welche Kriterien dabei den größten Stellenwert haben. In weiteren Untersuchungen könnte der BTQ-10 durch geschulte Beobachter und im Sinne eines Self-Assessement durch die Dozierenden angewandt werden. Somit wäre es möglich, die Abhängigkeit des Einflusses der zehn Kriterien auf die Gesamtbewertung von Unterrichtsqualität vom Beobachter zu spezifizieren und zu untersuchen, welche Kriterien für die Dozierenden selbst und für externe Beobachter den größten Einfluss auf die individuelle Lehrqualität haben.

6. LITERATURVERZEICHNIS

1. Blue AV, Griffith CH, 3rd, Wilson J, Sloan DA, Schwartz RW. Surgical teaching quality makes a difference. Am J Surg 1999; 177(1): 86-9.

2. Griffith CH, 3rd, Georgesen JC, Wilson JF. Six-year documentation of the association between excellent clinical teaching and improved students' examination performances. Acad Med 2000; 75(10 Suppl): S62-4.

3. Roop SA, Pangaro L. Effect of clinical teaching on student performance during a medicine clerkship. Am J Med 2001; 110(3): 205-9.

4. Stern DT, Williams BC, Gill A, Gruppen LD, Woolliscroft JO, Grum CM. Is there a relationship between attending physicians' and residents' teaching skills and students' examination scores? Acad Med 2000; 75(11): 1144-6.

5. Gibbs G, Coffey M. The Impact Of Training Of University Teachers on their Teaching Skills, their Approach to Teaching and the Approach to Learning of their Students. Active Learning in Higher Education 2004; 5(1): 87-100.

6. Schneider M, Mustafic M, editors. Gute Hochschullehre: Eine evidenzbasierte Orientierungshilfe. Berlin Heidelberg: Springer-Verlag; 2015.

7. Leach DC. Changing education to improve patient care. Qual Health Care 2001; 10 Suppl 2: ii54-8.

8. Fluit CR, Bolhuis S, Grol R, Laan R, Wensing M. Assessing the quality of clinical teachers: a systematic review of content and quality of questionnaires for assessing clinical teachers. J Gen Intern Med 2010; 25(12): 1337-45.

9. Leach DC, Philibert I. High-quality learning for high-quality health care: getting it right. JAMA 2006; 296(9): 1132-4.

10. Fluit C, Bolhuis S, Grol R, et al. Evaluation and feedback for effective clinical teaching in postgraduate medical education: validation of an assessment instrument incorporating the CanMEDS roles. Med Teach 2012; 34(11): 893-901.

11. Steinert Y, Mann K, Centeno A, et al. A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME Guide No. 8. Med Teach 2006; 28(6): 497-526.

12. Bierer SB, Hull AL. Examination of a clinical teaching effectiveness instrument used for summative faculty assessment. Eval Health Prof 2007; 30(4): 339-61.

13. McLean M, Cilliers F, Van Wyk JM. Faculty development: yesterday, today and tomorrow. Med Teach 2008; 30(6): 555-84.

14. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ 2003; 37(9): 830-7.

15. Snell L, Tallett S, Haist S, et al. A review of the evaluation of clinical teaching: new perspectives and challenges. Med Educ 2000; 34(10): 862-70.

16. Gjerde CL, Kokotailo P, Olson CA, Hla KM. A weekend program model for faculty development with primary care physicians. Fam Med 2004; 36 Suppl: S110-4.

17. Gunderman RB, Kang YP, Fraley RE, Williamson KB. Teaching the teachers. Radiology 2002; 222(3): 599-603.

18. Jafri W, Mumtaz K, Burdick WP, Morahan PS, Freeman R, Zehra T. Improving the teaching skills of residents as tutors/facilitators and addressing the shortage of faculty facilitators for PBL modules. BMC Med Educ 2007; 7: 34.

19. Morrison EH, Rucker L, Boker JR, et al. The effect of a 13-hour curriculum to improve residents' teaching skills: a randomized trial. Ann Intern Med 2004; 141(4): 257-63.

20. Orlander JD, Gupta M, Fincke BG, Manning ME, Hershman W. Co-teaching: a faculty development strategy. Med Educ 2000; 34(4): 257-65.

21. Sanci LA, Coffey CM, Veit FC, et al. Evaluation of the effectiveness of an educational intervention for general practitioners in adolescent health care: randomised controlled trial. BMJ 2000; 320(7229): 224-30.

22. Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? A review of the published instruments. J Gen Intern Med 2004; 19(9): 971-7.

23. Vaughan B. Developing a clinical teaching quality questionnaire for use in a university osteopathic pre-registration teaching program. BMC Med Educ 2015; 15: 70.

24. Schiekirka S, Feufel MA, Herrrmann-Lingen C, Raupach T. Evaluation in medical education: A topical review of target parameters, data collection tools and confounding factors. GMS German Medical Science 2015; 13: 1-19.

25. Dolmans DH, Wolfhagen HA, Gerver WJ, De Grave W, Scherpbier AJ. Providing physicians with feedback on how they supervise students during patient contacts. Med Teach 2004; 26(5): 409-14.

26. Donnelly MB, Woolliscroft JO. Evaluation of clinical instructors by third-year medical students. Acad Med 1989; 64(3): 159-64.

27. Irby D, Rakestraw P. Evaluating clinical teaching in medicine. J Med Educ 1981; 56(3): 181-6.

28. James PA, Kreiter CD, Shipengrover J, Crosson J. Identifying the attributes of instructional quality in ambulatory teaching sites: a validation study of the MedEd IQ. Fam Med 2002; 34(4): 268-73.

29. Shellenberger S, Mahan JM. A factor analytic study of teaching in off-campus general practice clerkships. Med Educ 1982; 16(3): 151-5.

30. Solomon DJ, Speer AJ, Rosebraugh CJ, DiPette DJ. The reliability of medical student ratings of clinical teaching. Eval Health Prof 1997; 20(3): 343-52.

31. Litzelman DK, Stratos GA, Marriott DJ, Skeff KM. Factorial validation of a widely disseminated educational framework for evaluating clinical teachers. Acad Med 1998; 73(6): 688-95.

32. Litzelman DK, Westmoreland GR, Skeff KM, Stratos GA. Factorial validation of an educational framework using residents' evaluations of clinician-educators. Acad Med 1999; 74(10 Suppl): S25-7.

33. Nation JG, Carmichael E, Fidler H, Violato C. The development of an instrument to assess clinical teaching with linkage to CanMEDS roles: A psychometric analysis. Med Teach 2011; 33(6): e290-6.

34. Beckman TJ, Lee MC, Rohren CH, Pankratz VS. Evaluating an instrument for the peer review of inpatient teaching. Med Teach 2003; 25(2): 131-5.

35. Donner-Banzhoff N, Merle H, Baum E, Basler HD. Feedback for general practice trainers: developing and testing a standardised instrument using the importance-quality-score method. Med Educ 2003; 37(9): 772-7.

36. Silber C, Novielli K, Paskin D, et al. Use of critical incidents to develop a rating form for resident evaluation of faculty teaching. Med Educ 2006; 40(12): 1201-8.

37. Steiner IP, Franc-Law J, Kelly KD, Rowe BH. Faculty evaluation by residents in an emergency medicine program: a new evaluation instrument. Acad Emerg Med 2000; 7(9): 1015-21.

38. Cohen R, MacRae H, Jamieson C. Teaching effectiveness of surgeons. Am J Surg 1996; 171(6): 612-4.

39. James PA, Osborne JW. A measure of medical instructional quality in ambulatory settings: the MedIQ. Fam Med 1999; 31(4): 263-9.

40. Zuberi RW, Bordage G, Norman GR. Validation of the SETOC instrument -- Student evaluation of teaching in outpatient clinics. Adv Health Sci Educ Theory Pract 2007; 12(1): 55-69.

41. Mullan P, Sullivan D, Dielman T. What are raters rating? Predicting medical student, pediatric resident, and faculty ratings of clinical teachers. Teaching and Learning in Medicine 1993; 5(4): 221-6.

42. Stalmeijer RE, Dolmans DH, Wolfhagen IH, Muijtjens AM, Scherpbier AJ. The development of an instrument for evaluating clinical teachers: involving stakeholders to determine content validity. Med Teach 2008; 30(8): e272-7.

43. Stalmeijer RE, Dolmans DH, Wolfhagen IH, Muijtjens AM, Scherpbier AJ. The Maastricht Clinical Teaching Questionnaire (MCTQ) as a valid and reliable instrument for the evaluation of clinical teachers. Acad Med 2010; 85(11): 1732-8.

44. Spickard A, 3rd, Corbett EC, Jr., Schorling JB. Improving residents' teaching skills and attitudes toward teaching. J Gen Intern Med 1996; 11(8): 475-80.

45. Williams BC, Litzelman DK, Babbott SF, Lubitz RM, Hofer TP. Validation of a global measure of faculty's clinical teaching performance. Acad Med 2002; 77(2): 177-80.

46. Bergen MR, Stratos GA, Berman J, Skeff KM. Comparison of clinical teaching by residents and attending physicians in inpatient and lecture settings. Teaching and Learning in Medicine: An International Journal 1993; 5(3): 149-57.

47. Hekelman FP, Vanek E, Kelly K, Alemagno S. Characteristics of family physicians' clinical teaching behaviors in the ambulatory setting: A descriptive study. Teaching and Learning in Medicine: An International Journal 1993; 5(1): 18-23.

48. McLeod PJ, James CA, Abrahamowicz M. Clinical tutor evaluation: a 5-year study by students on an in-patient service and residents in an ambulatory care clinic. Med Educ 1993; 27(1): 48-54.

49. Copeland HL, Hewson MG. Developing and testing an instrument to measure the effectiveness of clinical teaching in an academic medical center. Acad Med 2000; 75(2): 161-6.

50. Iblher P, Zupanic M, Härtel C, Heinze H, Schmucker P, Fischer M. Der Fragebogen "SFDP26-German": Ein verlässliches Instrument zur Evaluation des klinischen Unterrichts? Zeitschrift für medizinische Ausbildung 2011; 28(2).

51. Marriott DJ, Litzelman DK. Students' global assessments of clinical teachers: a reliable and valid measure of teaching effectiveness. Acad Med 1998; 73(10 Suppl): S72-4.

52. Skeff KM. Enhancing teaching effectiveness and vitality in the ambulatory setting. J Gen Intern Med 1988; 3(2 Suppl): S26-33.

53. Williams BC, Pillsbury MS, Kolars JC, Grum CM, Hayward R. Reliability of a global measure of faculty teaching performance. Journal of General Internal Medicine 1996; 12: 100.

54. Conigliaro RL, Stratton TD. Assessing the quality of clinical teaching: a preliminary study. Med Educ 2010; 44(4): 379-86.

55. Vu TR, Marriott DJ, Skeff KM, Stratos GA, Litzelman DK. Prioritizing areas for faculty development of clinical teachers by using student evaluations for evidence-based decisions. Acad Med 1997; 72(10 Suppl 1): S7-9.

56. Irby DM. Evaluating instructional scholarship in medicine. J Am Podiatr Med Assoc 1993; 83(6): 332-7.

57. Buchel TL, Edwards FD. Characteristics of effective clinical teachers. Fam Med 2005; 37(1): 30-5.

58. Morrison EH, Hitchcock MA, Harthill M, Boker JR, Masunaga H. The on-line Clinical Teaching Perception Inventory: a "snapshot" of medical teachers. Fam Med 2005; 37(1): 48-53.

59. Beckman TJ. Lessons learned from a peer review of bedside teaching. Acad Med 2004; 79(4): 343-6.

60. Beckman TJ, Mandrekar JN. The interpersonal, cognitive and efficiency domains of clinical teaching: construct validity of a multi-dimensional scale. Med Educ 2005; 39(12): 1221-9.

61. Elnicki DM, Cooper A. Medical students' perceptions of the elements of effective inpatient teaching by attending physicians and housestaff. J Gen Intern Med 2005; 20(7): 635-9.

62. Elnicki DM, Kolarik R, Bardella I. Third-year medical students' perceptions of effective teaching behaviors in a multidisciplinary ambulatory clerkship. Acad Med 2003; 78(8): 815-9.

63. Price DA, Mitchell CA. A model for clinical teaching and learning. Med Educ 1993; 27(1): 62-8.

64. Sutkin G, Wagner E, Harris I, Schiffer R. What makes a good clinical teacher in medicine? A review of the literature. Acad Med 2008; 83(5): 452-66.

65. Irby DM, Ramsey PG, Gillmore GM, Schaad D. Characteristics of effective clinical teachers of ambulatory care medicine. Acad Med 1991; 66(1): 54-5.

66. Marsh HW. Student´s Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness. In: Perry RP, Smart JC, eds. The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective: Springer Netherlands; 2007: 319-83.

67. Rindermann H. Lehrevaluation an Hochschulen: Schlussfolgerung aus Forschung und Anwendung für Hochschulunterricht und seine Evaluation. Zeitschrift für Evaluation 2003; 2: 233-56.

68. Marsh HW, Roche LA. Making students' evaluations of teaching effectiveness effective. American Psychologist 1997; 52(11): 1187-97.

69. Rindermann H. Die studentische Beurteilung von Lehrveranstaltungen - Forschungsstand und Implikationen. In: Spiel. C, ed. Evaluation universitärer Lehre - zwischen Qualitätsmanagement und Selbstzweck. Münster: Waxmann Verlag; 2001: 61-88.

70. Boerboom TB, Mainhard T, Dolmans DH, Scherpbier AJ, Van Beukelen P, Jaarsma AD. Evaluating clinical teachers with the Maastricht clinical teaching questionnaire: how much 'teacher' is in student ratings? Med Teach 2012; 34(4): 320-6.

71. Albanese MA, Schuldt SS, Case DE, Brown D. The validity of lecturer ratings by students and trained observers. Acad Med 1991; 66(1): 26-8.

72. Ramsbottom-Lucier MT, Gillmore GM, Irby DM, Ramsey PG. Evaluation of clinical teaching by general internal medicine faculty in outpatient and inpatient settings. Acad Med 1994; 69(2): 152-4.

73. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? J Gen Intern Med 2005; 20(12): 1159-64.

74. Himme A. Gütekriterien der Messung: Reliabilität, Validität und Generalisierbarkeit. In: Albers S, Klapper D, Konradt U, Walter A, Wolf J, eds. Methodik der empirischen Forschung. Wiesbaden: Gabler Verlag; 2007: 375-89.

75. Rammstedt B. Reliabilität, Validität, Objektivität. In: Wolf C, Best H, eds. Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften; 2010.

76. Bortz J, Döring N. Forschungsmethoden und Evaluation für Human -und Sozialwissenschaftler. Berlin, Heidelberg: Springer Verlag; 2006.

77. Bühner M. Einführung in die Test- und Fragebogenkonstruktion. München: Pearson Studium; 2011.

78. Moosbrugger H, Kelava A. Testtheorie und Fragebogenkonstruktion. Berlin, Heidelberg: Springer Verlag; 2011.

79. Häcker H, Leutner D, Amelang M. Standards für pädagogisches und psychologisches Testen: Hogrefe AG; 1998.

80. Schum T, Koss R, Yindra K, Nelson D. Students' and residents' ratings of teaching effectiveness in a department of pediatrics. Teaching and Learning in Medicine: An International Journal 1993; 5(3): 128-32.

81. McLeod PJ. Faculty Perspectives of a Valid and Reliable Clinical Tutor Evaluation Program. Eval Health Prof 1991; 14(3): 333-42.

82. Breckwoldt J, Lingemann C, Lingemann K. A theoretical framework to describe development of expertise in clinical teaching. The Association for Medical Education in Europe. Glasgow; 2010. p. 410-11.

83. Slavin RE. Quality, appropriateness, incentive, and time: A model of instructional effectiveness. International Journal of Educational Research 1994; 21(2): 141-57.

84. Brophy J. Teaching. 1999. http://www.ibe.unesco.org/en/services/online-materials/publications/educational-practices.html (accessed 19.06.2012)

85. Helmke A. Unterrichtsqualität erfassen, bewerten, verbessern. Seelze: Kallmeyer; 2005.

86. Meyer H. Was ist guter Unterricht? Berlin: Cornelson Verlag Scriptor 2004.

87. Edwards JR, Bagozzi RP. On the Nature and Direction of Relationships Between Constructs and Measures. Psychological Methods 2000; 5(2): 155-74.

88. Christophersen T. Usability im Online-Shopping: Entwicklung eines Fragebogeninstruments (ufos V2) unter Berücksichtiggung formativer und reflektiver Messmodelle. Kiel: Christian-Albrechts Universität; 2006.

89. Ullian JA, Bland CJ, DE. S. An alternative approach to defining the role of the clinical teacher. Acad Med 1994; 69(10): 832-8.

90. Christophersen T, Grape C. Die Erfassung latenter Konstrukte mit Hilfe formativer und reflektiver Messmodelle. In: Albers S, Klapper D, Konradt U, Walter A, Wolf J, eds. Methodik der empirischen Forschung. Wiesbaden: Gabler Verlag; 2007: 103-19.

91. Eberl M. Formative und reflektive Indikatoren im Forschungsprozess: Entscheidungsregeln und die Dominanz des refelektiven Modells. Schriften zur Empirischen Forschung und Quantitativen Unternehmensplanung. München: Ludwig-Maximilians-Universität; 2004. p. 1-44.

92. Huber F, Herrmann A, Meyer F, Vogel J, Vollhardt K. Kausalmodellierung mit Partial Least Squares. Wiesbaden Gabler Verlag; 2007.

93. Jarvis CB, MacKenzie SB, Podsakoff PM. A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research. Journal of Consumer Research 2003; 30(2): 199-218.

94. Bollen K, Lennox R. Conventional wisdom on measurement: a structural equation perspective. Psychological bulletin 1991; 110: 305-14.

95. Diamantopoulos A, Riefler P, Roth KP. Advancing formative measurement models. Journal of Business Research 2008; 61(12): 1203-18.

96. Schnell R, Hill PB, Esser E. Methoden der empirischen Sozialforschung. München: Oldenbourg Wissenschaftsverlag GmbH; 2005.

97. Lienert GA, Raatz U. Testaufbau und Testanalyse. Weinheim: Beltz Psychologie Verlags Union; 1998.

98. Diamantopoulos A. Formative indicators: Introduction to the special issue. Journal of Business Research 2008; 61(12): 1201-2.

99. Diamantopoulos A, Siguaw JA. Formative Versus Reflective Indicators in Organizational Measure Development: A Comparison and Empirical Illustration. British Journal of Management 2006; 17: 263–82.

100. Diamantopoulos A. Export performance measurement: reflective versus formative indicators. International Marketing Review 1999; 16(6): 444-57.

101. Schneider H. Nachweis und Behandlung von Multikollinearität. In: Albers S, Klapper D, Konradt U, Walter A, Wolf J, eds. Methodik der empirischen Forschung. Wiesbaden: Gabler Verlag; 2007: 183-98.

102. Rossiter JR. The C-OAR-SE procedure for scale development in marketing International Journal of Research in Marketing 2002; 16(4): 305-35.

103. Albers S, Hildebrandt L. Methodische Probleme bei der Erfolgsfaktorenforschung : Messfehler, formative versus reflektive Indikatoren und die Wahl des Strukturgleichungs-Modells. Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung 2006; 58: 2-33.

104. Temme D. Die Spezifikation und Identifikation formativer Messmodelle der Marketingforschung in Kovarianzstrukturanalysen. Marketing ZFP - Journal of Research and Management 2006; 28(3): 183-209.

105. Diamantopoulos A. The C-OAR-SE procedure for scale development in marketing: A comment. International Journal of Research in Marketing 2005; 22: 1-9.

106. MacKenzie SB, Podsakoff PM, Jarvis CB. The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. J Appl Psychol 2005; 90(4): 710-30.

107. Backhaus K, Erichson B, Weiber R. Fortgeschrittene multivariate Analysemethoden. Berlin: Springer Verlag; 2011.

108. Rohrmann B. Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. Zeitschrift für Sozialpsychologie 1978; 9: 222-45.

109. Carifio J, Perla RJ. Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. Journal of Social Sciences 2007; 3(3): 106-16.

110. Clason DL, Dormody TJ. Analyzing Data Measured by Individual Likert-Type Items. Journal of Agricultural Education 2010; 35(4): 31-5.

111. Forrest M, Andersen B. Ordinal scale and statistics in medical research. British Medical Journal 1986; 292: 537-8.

112. Harwell MR, Gatti G. Rescaling Ordinal Data to Interval Data in Educational Research Review of educational research 2001; 71(1): 105-31.

113. Jamieson S. Likert scales: how to (ab)use them. Medical Education 2004; 38(12): 1217-8.

114. Knapp TR. Treating Ordinal Scales as Interval Scales: An Attempt To Resolve the Controversy. Nursing research 1990; 39(2): 121-3.

115. Norman G. Likert scales, levels of measurement and the ‘‘laws’’ of statistics. Advances in health sciences education : theory and practice 2010; 15(5): 625-32.

116. Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaires. Journal of Rehabilitation Medicine 2001; 33(1): 47-8.

117. Kuzon WM, Urbanchek MG, McCabe S. The seven deadly sins of statistical analysis. Annals of Plastic Surgery 1996; 37(3): 265-72.

118. Schlarmann JG, Galatsch M. Regressionsmodelle für ordinale Zielvariablen GMS Medizinische Informatik, Biometrie und Epidemiologie 2014; 10(1): 1-9.

119. Schendera C. Regressionsanalyse mit SPSS. München: Oldenbourg Wissenschaftsverlag GmbH; 2008.

120. Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: John Wiley & Sons; 2000.

121. Rohrlack C. Logistische und Ordinale Regression. In: Albers S, Klapper D, Konradt U, Walter A, Wolf J, eds. Methodik der empirischen Forschung. Wiesbaden: Gabler Verlag; 2007: 199-215.

122. Backhaus K, Erichson B, Plinke W, Weiber R. Multivariate Analysemethoden - Eine anwendungsorientierte Einführung. Berlin: Springer Verlag; 2011.

123. Bender R, Ziegler A, Lange S. Logistische Regression. Deutsche Medizinische Wochenschrift 2007; 132: 33-5.

124. Best H, Wolf C. Logistische Regression. In: Wolf CB, H, ed. Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften; 2010.

125. Fahrmeir L, Kneib T, Lang S. Regression: Modelle, Methoden und Anwendungen. Heidelberg Dordrecht London New York: Springer; 2009.

126. Kühnel SM, Krebs D. Multinomiale und ordinale Regression. In: Wolf C, Best H, eds. Handbuch der sozialwissenschaftlichen Datenanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften; 2010.

127. Long JS. Regression Models for Categorical and Limited Dependent Variables. New York: Sage Publications; 1997.

128. Urban D. LOGIT-Analyse: Statistische Verfahren zur Analyse von Modellen mit qualitativen Response-Variablen. Stuttgart: Fischer; 1993.

129. McCullagh P. Regression models for ordinal data. Journal of the Royal Statistical Society 1980; 42(2): 109-42.

130. Urban D, Mayerl J. Regressionsanalyse: Theorie, Technik und Anwendung. Wiesbaden: VS Verlag für Sozialwissenschaften; 2008.

131. Bortz J, Schuster C. Statistik für Human- und Sozialwissenschaftler. Heidelberg: Springer Verlag; 2010.

132. McFadden D. The Measurement of urban travel demand. Journal of Public Economics 1974; 3(4): 303-28.

133. Cox DR, Snell EJ. The Analysis of Binary Data: Chapman and Hall; 1989.

134. Nagelkerke NJD. A note on the general definition of the coefficient of determination. Biometrika 1991; 78(3): 691-2.

135. Allison PD. Comparing Logit and Probit Coefficients Across Groups. Sociological Methods & Research 1999; 28(2): 186-208.

136. Williams R. Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients Across Groups. Sociological Methods & Research 2009; 37(4): 531-59.

137. Auspurg K, Hinz T. Gruppenvergleiche bei Regressionen mit binären abhängigen Variablen – Probleme und Fehleinschätzungen am Beispiel von Bildungschancen im Kohortenverlauf. Zeitschrift für Soziologie 2011; 40(1): 62-73.

138. Hoetker G. The use of logit and probit models in strategic management research: Critical issues. Strategic Management Journal 2007; 28(4): 331-43.

139. Keele L, Park DK. Difficult Choices: An Evaluation of Heterogeneous Choice Models. Paper for the 2004 Meeting of the American Political Science Association; 2006.

140. Mood C. Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. European Sociological Review 2010; 26(1): 67-82.

141. Williams R. Fitting heterogeneous choice models with oglm. The STATA Journal 2010; 10(4): 540-67.

142. Akaike H. Information theory and an extension of the maximum likelihood principle. Proceedings of the Second International Symposium on Information Theory Budapest: Akademiai Kiado 1973: 267-81.

143. Raftery AE. Bayesian Model Selection in Social Research. Sociological Methodolgy 1995; 25: 111-63.

144. Smith CA, Varkey AB, Evans AT, Reilly BM. Evaluating the performance of inpatient attending physicians: a new instrument for today's teaching hospitals. J Gen Intern Med 2004; 19(7): 766-71.

145. MacDonald PJ, Bass MJ. Characteristics of highly rated family practice preceptors. J Med Educ 1983; 58(11): 882-93.

146. Afonso NM, Cardozo LJ, Mascarenhas OA, Aranha AN, Shah C. Are anonymous evaluations a better assessment of faculty teaching performance? A comparative analysis of open and anonymous evaluation processes. Fam Med 2005; 37(1): 43-7.

7. ABBILDUNGSVERZEICHNIS

8. TABELLENVERZEICHNIS

9. EIDESSTATTLICHE VERSICHERUNG

„Ich, Anja Prescher, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: „Validierung eines empirisch basierten Beobachtungsinstruments für Unterrichtsqualität in der medizinischen Lehre" selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.

Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche in korrekter Zitierung (siehe „Uniform Requirements for Manuscripts (URM)“ des ICMJE -*www.**icmje**.org*) kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) entsprechen den URM (s.o) und werden von mir verantwortet.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Betreuer/in, angegeben sind. Sämtliche Publikationen, die aus dieser Dissertation hervorgegangen sind und bei denen ich Autor bin, entsprechen den URM (s.o) und werden von mir verantwortet.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§156,161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

20.04.2016

10. LEBENSLAUF

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

11. DANKSAGUNG

An dieser Stelle möchte ich mich bei allen Personen bedanken, die mich beim Zustandekommen dieser Arbeit unterstützt haben.

Ich bedanke mich bei Herrn Prof. Dr. Harm Peters für die Betreuung der Promotion.

Mein größter Dank geht an Dr. Jan Breckwoldt, der mir mit Ausdauer, Engagement, steter Zuversicht und einer Fülle an Ratschlägen zur Seite stand.

Ich möchte auch herzlich Herrn Christian Siggemann für die vielen Gespräche und Anregungen danken.

Ohne die Unterstützung und den Rückhalt meiner Familie und Freunde, allen voran meiner Eltern Christiane und Andreas Prescher und meinem Großvater Werner Gerwinski, wäre diese Arbeit nie möglich gewesen.