Aus der Klinik für Kardiologie und Pneumologie
(Prof. Dr. med. G. Hasenfuß)
der Medizinischen Fakultät der Universität Göttingen

Erprobung der Vorhersage-basierten Evaluationsmethode für das Instrument der Lernerfolgsevaluation

INAUGURAL-DISSERTATION

zur Erlangung des Doktorgrades
der Medizinischen Fakultät der
Georg-August-Universität zu Göttingen

vorgelegt von

Binia-Laureen Grebener, geb. Entgelmeier

aus

Flensburg

Göttingen 2021

Dekan:	Prof. Dr. med. W. Brück

Betreuungsausschuss

Betreuer/in:	Prof. Dr. med. T. Raupach, MME
Ko-Betreuer/in:	PD Dr. med. C. von der Brelie

Prüfungskommission

Referent/in:	Prof. Dr. med. T. Raupach, MME
Ko-Referent/in:	PD Dr. med. dent. Dr. rer. medic. P. Kanzow, MsC
Drittreferent/in:	Prof. Dr. med. R. Dressel

Datum der mündlichen Prüfung:	17.05.2022

Hiermit erkläre ich, die Dissertation mit dem Titel „Erprobung der Vorhersage-basierten Evaluationsmethode für das Instrument der Lernerfolgsevaluation“ eigenständig angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Göttingen, den 18.07.2021 ……

(Unterschrift)

Die Daten, auf denen die vorliegende Arbeit basiert, wurden teilweise publiziert:

Grebener BL, Barth J, Anders S, Beißbarth T, Raupach T (2021): A prediction-based method to estimate student learning outcome: Impact of response rate and gender differences on evaluation results. Med Teach 43, 524–530

.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

APG	*aggregated performance gain,* aggregierter Lernzuwachs
AUC	*area under the curve,* Fläche unter der Kurve
BAG	Bundesamt für Gesundheit
Chron. Bronchitis	Chronische Bronchitis
CSA	*comparative self-assessment,* vergleichende studentische Selbsteinschätzungen
Cut	*Cut-off*-Wert
DeGEval	Gesellschaft für Evaluation e.V.
EKG	Elektrokardiogramm
F11	Fach 11
IMPP	Institut für medizinische und pharmazeutische Prüfungsfragen
IPG	*individual performance gain,* individueller Lernzuwachs
MC-Fragen	*Multiple-Choice*-Fragen
MW	Mittelwert
NKLM	Nationaler Kompetenzbasierter Lernzielkatalog Medizin
NNE	*number needed to evaluate,* Anzahl der notwendigen Evaluationsteilnehmer
NPW	Negativ prädiktiver Wert
oAPG	*objective aggregated performance gain,* objektiver aggregierter Lernzuwachs
OBE	*outcome-based education,* lernergebnisorientierte Lehre
oIPG	*objective individual performance gain*, objektiver individueller Lernzuwachs
pAPG	*prediction-based aggregated performance gain,* Vorhersage-basierter aggregierter Lernzuwachs
pIPG	*prediction-based individual performance gain*, Vorhersage-basierter individueller Lernzuwachs
PPW	Positiv prädiktiver Wert
ROC-Analyse	*Receiver-operating-characteristics*-Analyse
sAPG	*subjective aggregated performance gain*, subjektiver aggregierter Lernzuwachs
Sens	Sensitivität
sIPG	*subjective individual performance gain*, subjektiver individueller Lernzuwachs
Spec	Spezifität
T1 bzw. T2	Zeitpunkt eins bzw. zwei
UMG	Universitätsmedizin Göttingen

1 Einleitung

1.1 Das Medizinstudium an der Universitätsmedizin Göttingen

1.1.1 Aufbau des Studiums

Das Studium der Humanmedizin an der Universitätsmedizin Göttingen (UMG) gliedert sich in einen vorklinischen und einen klinischen Studienabschnitt, gefolgt vom Praktischen Jahr. Der vorklinische Studienabschnitt umfasst vier Fachsemester, in denen vorwiegend naturwissenschaftliche Kenntnisse und Grundlagenfächer wie Anatomie, Physiologie und Biochemie gelehrt werden. Er schließt mit dem ersten Abschnitt der Ärztlichen Prüfung ab. Im Anschluss daran folgt der klinische Studienabschnitt. Dieser ist an der UMG modular strukturiert. Die insgesamt 21 Module erstrecken sich über einen Zeitraum von sechs Semestern und dauern jeweils zwischen zwei und sieben Wochen. Am Ende der drei Jahre findet der zweite Abschnitt der Ärztlichen Prüfung statt, gefolgt vom Praktischen Jahr und dem abschließenden dritten Abschnitt der Ärztlichen Prüfung. Jedes der 21 Module des klinischen Studienabschnittes hat einen thematischen Schwerpunkt. Die Schwerpunkte sind im Göttinger Lernzielkatalog in Form von Lernzielen für jedes Modul aufgelistet.

1.1.2 Der Göttinger Lernzielkatalog

Der Göttinger Lernzielkatalog „umfasst die Gesamtheit der prüfungsrelevanten Lehr- und Lerninhalte in der klinischen Ausbildung“ (Göttinger Lernzielkatalog 2017) und beinhaltet die zweite Ebene von Lernzielen des Nationalen Kompetenzbasierten Lernzielkatalogs Medizin (NKLM). Der NKLM stellt ein fächerübergreifendes Kerncurriculum für das Medizinstudium dar, welches vom Medizinischen Fakultätentag der Bundesrepublik Deutschland e.V. zusammen mit der Gesellschaft für Medizinische Ausbildung e.V. verfasst wurde (NKLM 2015). Der NKLM beinhaltet Lernziele, die als Kompetenzen formuliert sind. Kompetenzen werden von Epstein und Hundert wie folgt definiert:

> *„Professional competence is the habitual and judicious use of communication, knowledge, technical skills, clinical reasoning, emotions, values, and reflection in daily practice for the benefit of the individual and community being served“* (Epstein und Hundert 2002).

Die im NKLM festgelegten Kompetenzen sollen von den Studierenden am Ende des Studiums beherrscht werden und dienen den medizinischen Fakultäten als Leitfaden. Dabei lässt der NKLM den Fakultäten bewusst einen Freiraum, die Lernziele konkreter zu definieren und umzusetzen (NKLM 2015).

Der Göttinger Lernzielkatalog gliedert sich in vier Teile. In den Erläuterungen zum Aufbau des Lernzielkatalogs heißt es, dass sich „die Schwerpunktsetzung des Lernzielkatalogs […]

in der durchgeführten Lehre und den abgehaltenen Prüfungen widerspiegeln“ soll (Göttinger Lernzielkatalog 2017). Der erste Teil beinhaltet Symptome und Befunde, mit denen die Studierenden am Ende ihres Studiums vertraut sein sollen. Dadurch soll differentialdiagnostisches Denken gefördert werden. Der zweite Abschnitt ordnet Gesundheitsstörungen einem von drei Kompetenzniveaus zu, das von den Studierenden am Ende des Studiums in Bezug auf die verschiedenen Gesundheitsstörungen erreicht werden soll. Das Kompetenzniveau eins soll für epidemiologisch relevante und akut lebensbedrohliche Gesundheitsstörungen erreicht werden und erfordert ein tieferes Verständnis für die Ätiologie, Pathophysiologie, Diagnostik und Therapie der jeweiligen Erkrankungen. Für das Kompetenzniveau zwei sollen Studierende grundlegende Kenntnisse in Bezug auf die genannten Bereiche erwerben. Das Kompetenzniveau drei erfordert hingegen nur das Erkennen und Einordnen der entsprechenden Gesundheitsstörungen. Der dritte Abschnitt des Göttinger Lernzielkatalogs umfasst theoretische Kenntnisse sowie professionelle Einstellungen und Methoden. Zudem zählen ethische und rechtliche Grundlagen zu diesem Abschnitt. Im vierten Teil werden praktische Fertigkeiten aufgeführt, die zur späteren Ausübung des ärztlichen Berufs von Bedeutung sind (Göttinger Lernzielkatalog 2017).

1.2 *Outcome-based education*

Die Ausbildung von Ärzten[1] hat einen großen und wichtigen Stellenwert für die Gesellschaft. Um diesem gerecht zu werden, wird seit einigen Jahren der Fokus im Medizinstudium vermehrt auf *outcome-based education* (OBE) gelegt (Shumway und Harden 2003; Frank und Danoff 2007). Dieser Ansatz wird auch als „ergebnisorientiertes Denken“ bezeichnet (Harden 1999). Das Augenmerk liegt hierbei auf dem Resultat, also den Kompetenzen, die am Ende eines bestimmten Kurses oder Studienabschnittes erreicht werden sollen (Harden 2007a), wobei dieses Resultat den Weg dorthin bestimmt (Shumway und Harden 2003). Dabei definieren die zuvor festgelegten Lernziele, was gelehrt und geprüft wird (Harden 1999). Das Konzept OBE erfordert dazu eine enge Übereinstimmung von Lernzielen, Lehrinhalten und -methoden, Lernmethoden und Prüfungen und trägt hierdurch zur Sicherung der Lehrqualität bei (Harden 2007b).

Spady definiert diese Lernziele wie folgt: *„Outcomes are clear learning results that we want students to demonstrate at the end of significant learning experiences“* (Spady 1994). Die Studierenden sollen demnach Kenntnisse und Fähigkeiten, im Sinne von Kompetenzen, erwerben. Die Lernziele geben dabei das anvisierte Ziel in Bezug auf diese Kompetenzen vor. Lernziele lassen sich nach der Lernzieltaxonomie von Bloom et al. (1956) einer von drei Dimensionen zuordnen.

[1] Aus Gründen der besseren Lesbarkeit werden in der vorliegenden Arbeit die männlichen oder neutralen Formen personenbezogener Hauptwörter verwendet. Diese beziehen sich jedoch grundsätzlich in gleichem Maße auf alle Geschlechter und implizieren keine Benachteiligungen.

Es werden „kognitive, psychomotorische und affektive“ Lernziele unterschieden (Thomas et al. 2016).

Dabei sollen die Lernziele laut Spady (1988) die Ergebnisse widerspiegeln, die von den Studierenden erzielt werden sollen und auf dieser Grundlage ein Curriculum erstellt werden, anstatt Lernziele zu definieren, die zu einem bereits bestehenden Curriculum passen. Harden (1999) schlägt das Konzept OBE deshalb als Grundlage für die Curriculums-Planung und Evaluation vor. Lernziele müssen klar definiert werden, denn diese stellen die Grundlage für alle weiteren Entscheidungen in Bezug auf das Curriculum dar und sollen daher eine Schlüsselposition erhalten (Harden 1999).

In diesem Zusammenhang wird die Rolle von Evaluationen als Teil der zielgerichteten Entwicklung von Lehrcurricula betrachtet. Im sechsstufigen Konzept zur Curricularentwicklung von Thomas et al. (2016) erfolgen zunächst die Problemidentifikation und die allgemeine Bedarfsanalyse (Schritt eins). Im Anschluss werden die Bedürfnisse der Zielgruppe identifiziert (Schritt zwei) und Lernziele definiert (Schritt drei). In einem vierten Schritt werden zu den Lernzielen passende Lehrformate gewählt und Ausbildungsstrategien erarbeitet. Schritt fünf beinhaltet die Implementierung des Curriculums, gefolgt von der Evaluation im sechsten Schritt. In diesem Zirkel laufen alle Schritte parallel ab und bilden einen dynamischen Prozess. Dabei kann die Evaluation einen direkten Einfluss auf die anderen Schritte haben und stellt keinesfalls nur den letzten Schritt dar (Thomas et al. 2016). Auch für die Abstimmung der in der vorliegenden Studie genutzten Evaluationsaussagen mit der objektiven Prüfung und den Lernzielen aus dem Göttinger Lernzielkatalog wird das beschriebene Konzept (Thomas et al. 2016) genutzt.

Evaluationsinstrumente, die im Sinne der OBE auf definierten Lernzielen basieren, können somit bei Nicht-Erreichen der vorgegebenen Lernziele auf Defizite innerhalb des Curriculums schließen lassen (Harden 1999). Auch Gibson et al. (2008) empfehlen die Ausrichtung von Evaluationsmethoden an Lernzielen.

Die in diesem Kapitel dargestellten Grundlagen der OBE begründen die Wichtigkeit der Nutzung spezifischer Lernziele zur Evaluation der Lehre und stellen somit die Basis für die vorliegende Studie dar. Die Ausrichtung von Evaluationsinstrumenten an spezifischen Lernzielen wird in Kapitel 1.3.5 anhand des Beispiels an der UMG ausführlich beschrieben.

1.3 Evaluation

1.3.1 Eine Definition

Laut dem Schweizer Bundesamt für Gesundheit (BAG) umfasst eine Evaluation

> „die kritische, analytische Interpretation gewonnener Informationen, das Ziehen von Schlussfolgerungen daraus und, letztlich, die Beurteilung und/oder Bewertung eines Projektes oder einer Sachlage mit dem Ziel, diese zu verbessern“ (BAG 1997).

Es wird deutlich, dass eine Evaluation ein mehrstufiger Prozess ist, der neben der reinen Beschreibung und Messung der bestehenden Situation auch der Bewertung und Optimierung dieser dient und dazu beitragen soll, Defizite aufzudecken (Rindermann 2003). Laut der Gesellschaft für Evaluation e.V. (DeGEval) existieren vier Merkmale, die die grundlegenden Voraussetzungen für eine gute Evaluation darstellen: „Nützlichkeit", „Durchführbarkeit", „Fairness" und „Genauigkeit" (DeGEval 2016).

Im Folgenden wird der Fokus auf Evaluationen im Hochschulkontext und v. a. auf das Medizinstudium gerichtet. Es wird ein Überblick über studentische Bewertungen zur Evaluation der Lehre gegeben und der Nutzen studentischer Selbsteinschätzungen erläutert. Darüber hinaus soll die Vorhersage-basierte Methode zur Kursevaluation vorgestellt werden. Des Weiteren wird das bestehende Evaluationssystem an der UMG dargelegt und eine Einordnung der vorliegenden Arbeit zur Weiterentwicklung dieses Evaluationsinstrumentes vorgenommen.

1.3.2 Evaluation in der Hochschullehre

Evaluationen haben einen besonderen Stellenwert in der Hochschullehre, wobei das Ziel aller Bemühungen zur Evaluation der Lehre die Verbesserung der Lehrqualität ist. Es gibt viele verschiedene Arten der Durchführung von Evaluationen. In den 1990er Jahren etablierten sich an deutschsprachigen Hochschulen im Rahmen der Einführung von Evaluationen neben der internen (z. B. in Form von „Lehrberichten") und externen Evaluation (z. B. durch Experten) auch studentische Lehrveranstaltungsevaluationen (Rindermann 2003). In den folgenden Abschnitten wird der Fokus auf die studentischen Lehrveranstaltungsevaluationen gelegt, da das Verständnis dieser eine wichtige Voraussetzung zur Einordnung der vorliegenden Studie darstellt. Zudem basieren die meisten Lehrevaluationen im Medizinstudium auf studentischen Bewertungen (Woloschuk et al. 2011).

Studentische Evaluationen der Lehre dienen verschiedenen Zwecken. Dabei tragen sie beispielsweise zur Verbesserung der Lehre bei (Benton und Cashin 2011). Zudem werden Personalentscheidungen auf der Basis studentischer Evaluationen getroffen (Linse 2017) und die Entwicklung von Curricula wird unter anderem durch Evaluationen gelenkt (Shumway und Harden 2003). Des Weiteren dienen sie als Feedback für die Lehrenden und Organisatoren (McKeachie 1997; Kogan und Shea 2007). An einigen medizinischen Fakultäten in Deutschland werden studentische Bewertungen zudem als Basis für die Leistungsorientierte Mittelvergabe für Lehre herangezogen (Schiekirka-Schwake et al. 2019). Der Wissenschaftsrat erklärte in seinen Empfehlungen zur Qualitätsverbesserung von Lehre und Studium, dass Lehrevaluationen auf Basis von studentischen Bewertungen ein wichtiges, jedoch bisher kaum standardisiertes Instrument zur Erfassung der Lehrqualität darstellen (Wissenschaftsrat 2008). Darüber hinaus werden sie zu selten genutzt, um systematisch zur Optimierung der Lehre beizutragen (Wissenschaftsrat 2008). Germain und Scandura (2005)

bemängeln zudem, dass es an Kriterien zur Definition von erfolgreicher Lehre fehlt, die als Grundlage für die Evaluationen dienen. Beispielsweise definieren Dozierende und Studierende den Lernerfolg unterschiedlich und haben verschiedene Vorstellungen bezüglich der Messung dieses Lernerfolgs (Schiekirka et al. 2012). Billings-Gagliardi et al. (2004) führten eine Studie durch, in der Studierende während des Ausfüllens eines Evaluationsbogens laut ihre Gedanken äußern sollten. Dabei konnten sie zeigen, dass das Verständnis der genutzten Begriffe und Konzepte hinter den Evaluationen sowohl zwischen den einzelnen Studierenden, als auch zwischen Studierenden und Lehrenden differiert.

Die Tatsache, dass erfolgreiche Lehre sich nicht in einem einzelnen Kriterium zusammenfassen lässt, ist laut Marsh (1984) ein Grund dafür, weshalb die Validierung von studentischen Bewertungen schwierig ist und sich nicht durch ein einzelnes Argument bestätigen oder widerlegen lässt. Außerdem bestehen teilweise Fehlinformationen und Missverständnisse bezüglich der Nutzung studentischer Bewertungen (Theall und Franklin 2001; Linse 2017). Grund hierfür sind beispielsweise mangelnde Kenntnisse oder fehlende Akzeptanz der bestehenden Literatur zur Reliabilität und Validität studentischer Bewertungen von Seiten der Evaluierenden (Theall und Franklin 2001). So konnten Franklin und Theall (1989) in einer Studie zeigen, dass ein großer Anteil der Evaluierenden keine ausreichenden Kenntnisse der statistischen Grundlagen oder fehlendes Wissen über die Literatur zur Interpretation der Ergebnisse aus den studentischen Bewertungen aufwies. Geringe Kenntnisse der Evaluierenden zeigten dabei einen Zusammenhang mit einer negativen Einstellung gegenüber studentischen Bewertungen und Evaluationen (Franklin und Theall 1989).

1.3.2.1 Reliabilität und Validität studentischer Evaluationen

Studentische Evaluationen der Lehre sind eines der am meisten diskutierten Instrumente zur Lehrevaluation (Gravestock und Gregor-Greenleaf 2008). Es gibt immer wieder kontroverse Diskussionen darüber, ob studentische Bewertungen der Lehre valide und reliabel sind. Damit geht die Frage einher, ob sie ein geeignetes Instrument zur adäquaten Beurteilung der Lehrqualität darstellen. In den letzten Jahrzehnten befassten sich viele Autoren mit dieser Fragestellung (Greenwald 1997), sodass in der Literatur mehr als 2000 Arbeiten hierzu existieren (Murray 2005). Um die Qualität der Lehre angemessen zu beurteilen, ist das Vorhandensein reliabler und valider Kriterien eine zwingend notwendige Vorrausetzung. Die Validität beschreibt die Tatsache, ob ein Messinstrument tatsächlich das misst, was es beabsichtigt zu messen (Tavakol und Dennick 2011). In Bezug auf studentische Bewertungen der Lehre zur Evaluation der Lehrqualität gibt die Validität also an, ob ein Instrument tatsächlich die Lehrqualität misst. Die Reliabilität – auch Zuverlässigkeit – eines Evaluationsinstrumentes beschreibt, ob bei wiederholter Durchführung das gleiche Ergebnis erzielt wird. Im Hinblick auf studentische Bewertungen der Lehre ist hier häufig die *Inter-rater*-Reliabilität oder auch die interne Konsistenz gemeint (Benton und Cashin 2011; Kogan und Shea 2007). Viele Autoren bestätigen die Validität und Reliabilität studentischer

Bewertungen zur Evaluation der Lehre (Marsh 1987; d'Apollonia und Abrami 1997; McKeachie 1997; Wachtel 1998; Theall und Franklin 2001; Kogan und Shea 2007; Benton und Cashin 2011). Nichtsdestotrotz gibt es weiterhin Kritiker in Bezug auf den Nutzen studentischer Evaluationen (Wachtel 1998; Mason et al. 2002).

Ein Argument, das die Validität und Reliabilität studentischer Bewertungen in Frage stellt, ist die Beeinflussung durch verschiedene (Stör-)Faktoren, auch *bias* genannt. Der Ausdruck *bias* bezeichnet dabei ein Merkmal (der Dozierenden, der Studierenden oder des Kurses), welches das Ergebnis der Evaluation verändert, ohne dabei eine Aussage über die Qualität der Lehre zu treffen (Centra 2003). Spiel und Gössler definieren diese Faktoren als „Variablen, die in keinem systematischen Zusammenhang mit Lehrqualität stehen" (Spiel und Gössler 2000). In zahlreichen Studien konnten Faktoren identifiziert werden, die potenziell zu einer Beeinflussung der studentischen Bewertungen führen können. Gravestock und Gregor-Greenleaf (2008) ordnen diese Variablen vier Kategorien zu: „Organisatorischen Faktoren", „Kurs-Charakteristika", „Dozierenden-Charakteristika" sowie „Studierenden-Charakteristika". Zu den organisatorischen Faktoren zählen beispielweise der Zeitpunkt der Evaluation oder die Anwesenheit der Dozierenden. Zu den Studierenden-Charakteristika zählen unter anderem das Geschlecht, die Teilnahmerate, die Noten oder die Motivation der Studierenden (Gravestock und Gregor-Greenleaf 2008). Mehrere Autoren bestätigten, dass ein vorbestehendes Interesse der Studierenden für das Thema eines Kurses mit einer besseren Bewertung des Kurses assoziiert ist (Marsh 1987; Spiel und Gössler 2000; Berger und Schleußner 2003; Schiekirka und Raupach 2015), wobei nicht immer klar differenziert werden kann, ob das Interesse schon vor dem Kurs bestand oder sich im Laufe des Kurses entwickelt hat (Marsh und Roche 1997). Dybowski et al. (2017) beschrieben passend dazu, dass ein vorbestehendes Interesse von Studierenden für das Thema des Kurses, im Sinne eines *bias*, die studentischen Bewertungen der Lehrqualität beeinflusst. Bezüglich der erzielten oder erwarteten Noten, die Studierende in einem Kurs erhalten, wiesen einige Studien einen Zusammenhang mit der Bewertung des Kurses auf (Greenwald und Gillmore 1997). Auch Howard und Maxwell (1980) bestätigten den Zusammenhang zwischen den erwarteten Noten und der studentischen Zufriedenheit mit dem Kurs. Sie interpretierten diesen Zusammenhang jedoch als erwartete und erklärbare Variable, die auf studentischen Eigenschaften, wie etwa der Motivation, beruht. Zudem können Prüfungen, im Sinne einer negativen Erfahrung, einen starken Einfluss auf die Gesamtbewertungen eines Kurses haben (Woloschuk et al. 2011). Zu den dozentenspezifischen Charakteristika zählen Gravestock und Gregor-Greenleaf (2008) etwa das Geschlecht, den Ruf oder das Alter des Dozierenden. Bezüglich des Geschlechts können diskrepante Ergebnisse hinsichtlich des Einflusses auf die studentischen Bewertungen beobachtet werden (Marsh und Roche 1997; Aleamoni 1999). Ein Zusammenhang zwischen dem Ruf der Dozierenden und den studentischen Bewertungen konnte ebenfalls in einigen Studien festgestellt werden (Griffin 2001). Bezüglich der Beeinflussung durch Kurs-Charakteristika, wie etwa der Anzahl der Kursteilnehmer, gibt es unterschiedliche Meinungen in der Literatur (Marsh und Roche 1997;

Aleamoni 1999), wobei die Mehrheit der Studien eher einen inversen Zusammenhang zwischen der Kursgröße und der Bewertung des Kurses zeigt (Marsh und Roche 1997; Aleamoni 1999).

Viele Evaluationsinstrumente basieren dabei auf allgemeinen Bewertungen der Lehre durch globale Items (z. B. in Form von Schulnoten). Einige Autoren kritisieren, dass aber genau diese globalen Items zur Gesamtbewertung eines Kurses, im Gegensatz zu spezifischen Items, besonders anfällig für eine Beeinflussung durch die genannten Faktoren sind (Marsh und Roche 1997; Aleamoni 1999; Schiekirka und Raupach 2015; Schiekirka et al. 2015). Zudem können Marsh und Roche (1997) zufolge Globalbewertungen die verschiedenen Dimensionen der Lehre nicht hinreichend abbilden. Aus diesem Grund hängt die Validität und die Nützlichkeit studentischer Bewertungen der Lehre eng mit den Eigenschaften der genutzten Items zusammen (Marsh und Roche 1997). Auch muss das dem Evaluationsinstrument zugrunde liegende Konstrukt von „guter Lehre" eindeutig definiert sein und durch die genutzten Methoden valide gemessen werden (Raupach et al. 2012). Nur durch eine klare Definition von „guter Lehre" können geeignete Evaluationsinstrumente entwickelt werden, die ebendieses zugrundeliegende Konstrukt messen (Schiekirka-Schwake et al. 2019). Des Weiteren erfassen viele Evaluationsmethoden lediglich einzelne Aspekte der Lehre (Gibson et al. 2008). Gibson et al. (2008) beschrieben hingegen vier Dimensionen zur Erfassung der Lehrqualität, welche „Strukturen", „Prozesse", „Charakteristika der Dozierenden" sowie das „Ergebnis der Lehre" umfassen. Die Dimension „Strukturen" beinhaltet dabei beispielsweise den Aufbau des Studiums sowie vorhandene Materialien, während die Dimension „Prozesse" die Lehr- und Lernumgebung, Interaktionen und die Organisation abdeckt. Durch die Dimension „Charakteristika der Dozierenden" werden unter anderem die didaktischen Qualifikationen und durch das „Ergebnis der Lehre" beispielsweise der Lernerfolg erfasst (Gibson et al. 2008). In traditionellen Evaluationsinstrumenten wird vorwiegend die studentische Zufriedenheit mit dem Kurs und der Organisation abgebildet (Braun und Leidner 2009). Stattdessen sollte die Lehrqualität in Form von Lernerfolg, der eigentlich gemessen werden sollte, evaluiert werden (Rindermann und Schofield 2001; Gibson et al. 2008; Braun und Leidner 2009). Schiekirka et al. (2012) konnten zeigen, dass es Studierenden schwer fällt, Globalbewertungen eines Kurses abzugeben, da das persönliche Interesse hinsichtlich des Themas oder auch der Schwierigkeitsgrad der Abschlussprüfung diese globalen Bewertungen beeinflussen.

Es wird deutlich, dass die meisten Autoren die Reliabilität und Validität studentischer Bewertungen und deren Nutzen zur Bewertung der Lehre unterstützen (Wachtel 1998). Trotzdem besteht bisher keine vollständige Einigkeit über den Einfluss bestimmter Faktoren auf die studentischen Bewertungen. Marsh (1987) fasste zusammen, dass diese Faktoren, wenn überhaupt, einen kleinen Effekt auf die Bewertungen der Studierenden haben.

> *„These findings indicate that class-average student ratings are: 1) multidimensional; 2) reliable and stable; 3) primarily a function of the instructor who teaches a course rather than of the course that is*

taught; 4) relatively valid against a variety of indicators of effective teaching; 5) relatively unaffected by a variety of variables hypothesized as potential biases; and 6) seen to be useful by faculty as feedback about their teaching, by students for use in course selection, and by administrators for use in personnel decisions" (Marsh 1987).

Nach dieser Einschätzung sind studentische Evaluationen ein weitestgehend valides und reliables Instrument zur Evaluation der Lehre (Benton und Cashin 2011). Sie stellen einen wichtigen Bestandteil zur Erfassung der Lehrqualität dar, der jedoch nicht für sich alleine betrachtet werden sollte (Berk 2013). Stattdessen sollten studentische Bewertungen mit anderen Methoden zur Evaluation der Lehrqualität kombiniert werden (Benton und Cashin 2011).

Es scheint zudem ein Mangel alternativer und zugleich reliabler Formen der Hochschulevaluation zu bestehen, sodass trotz bestehender Zweifel weiterhin studentische Bewertungen zur Evaluation der Lehre genutzt werden (Germain und Scandura 2005). In einer bundesweiten Studie gaben alle teilnehmenden, medizinischen Fakultäten an, studentische Evaluationen zu nutzen, während Lehrende und externe Gutachter seltener zur Evaluation der Lehre herangezogen werden (Schiekirka-Schwake et al. 2019). Studierende stellen diejenige Zielgruppe dar, die unmittelbar von der Qualität der Lehre betroffen ist. Diese Tatsache unterstützt den Nutzen studentischer Bewertungen zur Evaluation. Ihre Einschätzungen und Bewertungen hierzu werden deshalb immer einen wichtigen Nutzen haben (McKeachie 1997; Theall und Franklin 2001). Allerdings weisen Theall und Franklin (2001) darauf hin, dass Studierende nicht in der Lage sind, die Gesamtheit der Eigenschaften der Lehre zu evaluieren. Sie sind nicht dazu qualifiziert, den Wissensstand der Dozierenden für das Thema des Kurses zu beurteilen, jedoch sehr wohl in der Lage, ihren eigenen Lernzuwachs in Bezug auf den ursprünglichen Wissensstand einzuschätzen oder ihre Zufriedenheit mit dem Kurs auszudrücken (Theall und Franklin 2001). Zudem kann von Studierenden nicht erwartet werden, dass sie die Aktualität des ausgegebenen Lehrmaterials bewerten oder die Kompetenz der Dozierenden bezüglich des Themas einschätzen können (Seldin 1989). Es muss also sehr genau abgewogen werden, welche Art von Evaluation in der Lehre auf der Basis studentischer Bewertungen sinnvoll ist. Laut Braun und Leidner (2009) ist es sinnvoll den Nutzen, den Studierenden durch die stattgefundene Lehre erhalten haben, zu evaluieren, anstelle ihrer Zufriedenheit damit. Aus diesem Grund sollte der Fokus vermehrt auf ergebnis- und kompetenzorientierte Evaluationen gelegt werden (Braun und Leidner 2009).

1.3.2.2 Evaluation im Medizinstudium

Seit der Einführung der Neuen Approbationsordnung für Ärzte 2002 gibt es Neuerungen bezüglich der Evaluation im Medizinstudium. Neben der Betonung der Wichtigkeit von qualitativ hochwertiger Lehre schreibt die Reform seither die regelmäßige Evaluation von Lehrveranstaltungen zur Erfassung der Lehrqualität vor (Güntert et al. 2003). Jedoch zeigte sich in einer bundesweiten Befragung, dass nur an 21 % der teilnehmenden Fakultäten eine

fakultätseigene Evaluationsordnung existiert, während sich der Großteil der Fakultäten auf die Ordnungen der Universitäten beruft (Schiekirka-Schwake et al. 2019). Die Evaluation im Medizinstudium unterscheidet sich jedoch im Hinblick auf die zu erfassende Lehre in mehreren Punkten von Evaluationen in anderen Studiengängen. Einige Lehrformate existieren in dieser Form nur im Medizinstudium, wie etwa der Unterricht am Krankenbett (Schiekirka et al. 2015). Auch der Einsatz von Schauspielpatienten zur Verbesserung kommunikativer Fähigkeiten der Studierenden findet immer mehr Einzug in die Lehre (Simmenroth et al. 2007). Diese Lehrformate fallen unter anderem in den Bereich der klinischen Lehre, die ein besonderes Merkmal im Gegensatz zu anderen Studienfächern darstellt. Die klinische Lehre ist meist durch kleinere, fest eingeteilte Gruppen gekennzeichnet, die durch eine zugeteilte Lehrperson für einen bestimmten Zeitraum eng betreut werden (Shumway und Harden 2003; Kogan und Shea 2007). Die Dauer der unterschiedlichen Kurse und Module schwankt dabei im Vergleich zu anderen Studiengängen stark und Medizinstudierende haben, verglichen mit Studierenden anderer Fächer, wenige Freiheiten in der Wahl der Kurse oder Dozierenden (Kogan und Shea 2007). Zudem werden Vorlesungen innerhalb eines Kurses häufiger als in anderen Studienfächern von mehreren verschiedenen Dozierenden gehalten (Kogan und Shea 2007; McOwen et al. 2009). Dadurch erlaubt eine globale Bewertung keine adäquate Beurteilung einzelner Dozenten. Die genannten Unterschiede zwischen der Lehre an medizinischen Fakultäten und der Lehre in anderen Studiengängen betreffen alle vier Dimensionen der Lehre (Gibson et al. 2008). Solche Besonderheiten in der Lehre stellen auch an die genutzten Evaluationsinstrumente spezielle Ansprüche. *„Specific evaluation tools need to reflect aspects that are specific to medical education"* (Müller et al. 2017). Dadurch kann der bestmögliche Nutzen aus den Evaluationen gewonnen werden. Es ist jedoch nicht jede Form der Evaluation hierzu geeignet, denn die Wahl eines Evaluationsverfahrens hängt größtenteils davon ab, welchem Zweck eine Evaluation dient (BAG 1997).

Um zu klären, welche Methoden zur Evaluation der Lehre im Medizinstudium geeignet sind, muss deshalb zunächst definiert werden, welches Konstrukt gemessen werden soll. Die Lehrqualität ist dabei weit mehr als ein eindimensionales Konstrukt (Gibson et al. 2008). In Kapitel 1.3.2.1 wurden bereits die von Gibson et al. (2008) beschriebenen vier Dimensionen („Strukturen", „Prozesse", „Charakteristika der Dozierenden", „Ergebnis der Lehre") zur Evaluation der Lehrqualität an medizinischen Fakultäten vorgestellt. Auch Herzig et al. (2007) definierten drei ähnliche Qualitätsebenen der Lehre: „Prozess-, Struktur- und Ergebnisqualität", mit einem besonderen Gewicht auf Letzterer. Bislang liegt der Fokus von Lehrveranstaltungsevaluationen in der medizinischen Lehre laut Kogan und Shea (2007) zumeist auf den Prozessen. Viele der an Universitäten genutzten Evaluationsmethoden basieren dabei auf allgemeinen Bewertungen. Die Evaluationsbögen verlangen nach globalen Einschätzungen zur Bewertung der Lehre und beziehen sich vorwiegend auf die Struktur- und Prozessqualität, während die Ergebnisqualität wenig Beachtung findet (Raupach et al. 2012). Auch Schiekirka-Schwake et al. (2019) berichteten, dass an deutschen medizinischen

Fakultäten vorwiegend Strukturen, Prozesse und der Gesamteindruck evaluiert werden. Für die Bewertung von Strukturen, Prozessen und individuellen Dozierenden existieren geeignete Erhebungsinstrumente, während sich die Erhebung der Lehrqualität in Form des Lehrergebnisses bisweilen schwierig gestaltet (Schiekirka et al. 2015). Diese Instrumente sind häufig allgemein gefasst und nicht an spezifische Lernziele gekoppelt, sodass sie wenig Aufschluss darüber geben, ob und inwieweit spezifische Ziele innerhalb eines Kurses erreicht werden (Raupach et al. 2011).

Wie im vorangegangenen Kapitel angedeutet bietet es sich an, den Fokus vermehrt auf die Ergebnisqualität zu legen. Zur Erfassung der Ergebnisqualität werden an einigen medizinischen Fakultäten in Deutschland die Ergebnisse des schriftlichen Teils des Zweiten Staatsexamens genutzt (Schiekirka-Schwake et al. 2019). Jedoch sind Prüfungsergebnisse immer nur Momentaufnahmen des Leistungsstandes der Studierenden und spiegeln damit nur den aktuellen Leistungsstand zu einem bestimmten Zeitpunkt wider, ohne den Lernerfolg über die Lehrveranstaltung hinweg zu betrachten (Schiekirka et al. 2015). Der Leistungsstand der Studierenden vor und nach einem Kurs ist zudem sehr variabel, sodass das Ergebnis einer Prüfung am Ende eines Kurses keine globalen Schlussfolgerungen über den tatsächlichen Zugewinn an Wissen, Kompetenzen oder Fähigkeiten einzelner Studierender zulässt (Raupach et al. 2011). Es sind stattdessen gepaarte Vergleiche einzelner Studierender vor und nach einem Kurs notwendig um zu erörtern, in welchen Bereichen die Lehre innerhalb des Kurses zu einem guten Zuwachs an Wissen oder Kompetenzen geführt hat und wo es Verbesserungsbedarf gibt (Raupach et al. 2011). Zudem werden an medizinischen Fakultäten in Deutschland schriftliche Prüfungen zu einem Großteil im *Multiple-Choice*-Format durchgeführt (Möltner et al. 2010). Auch in den schriftlichen Staatsexamensprüfungen wird dieses Aufgabenformat verwendet (IMPP – Zusammenstellung der Prüfungsinhalte für den Zweiten Abschnitt der Ärztlichen Prüfung). Allerdings besteht bei diesen geschlossenen Aufgaben das Problem, dass durch eine bestimmte Ratewahrscheinlichkeit (im Falle von fünf Antwortmöglichkeiten entspricht diese 20 %) die Frage auch dann zufällig richtig beantwortet werden kann, wenn das Ergebnis nicht sicher gekannt wird (Möltner et al. 2006). Ein weiteres grundlegendes Problem von *Multiple-Choice*-Fragen (MC-Fragen) ist zudem die eingeschränkte Fähigkeit zur Messung von klinischen Kompetenzen. So können MC-Fragen lediglich das erste und allenfalls das zweite Niveau der Miller-Pyramide (Miller 1990) abdecken (Wass et al. 2001). Das erste Niveau entspricht dabei dem Abrufen von Faktenwissen, während das zweite Niveau die Anwendung dieses Faktenwissens zur Problemlösung und Entscheidungsfindung darstellt (Miller 1990). Zur Überprüfung dieser Basis-Kompetenzen können Prüfungen basierend auf MC-Fragen eingesetzt werden, allerdings können diese keine Kompetenzen aus einem höheren Niveau der Miller-Pyramide, z. B. die praktische Durchführung der gelernten Kompetenzen an Simulationspatienten oder in realen Patientenfällen, abbilden (Wass et al. 2001).

Bei der Betrachtung von Prüfungen muss außerdem zwischen summativen und formativen Prüfungen unterschieden werden. Summative Prüfungen bewerten den Leistungsstand von

Studierenden am Ende eines Lehrabschnitts (Kibble 2017), stellen den Lernfortschritt sicher (Anziani et al. 2008) und dienen der Notenvergabe oder der Erstellung von Ranglisten (Roediger und Karpicke 2006). Durch die Notenvergabe stellen sie einen Anreiz zur Prüfungsvorbereitung der Studierenden auf diese Prüfung dar und führen hierdurch zu einer Änderung im Lernverhalten der Studierenden (Sennhenn-Kirchner et al. 2018). Formative Prüfungen verfolgen hingegen das Ziel der Verbesserung des Lernprozesses durch die Erteilung von Feedback (Kibble 2017). Durch dieses Feedback erhalten die Studierenden eine Rückmeldung zu ihrer Leistung, um Probleme identifizieren und die eigenen Fähigkeiten verbessern zu können (Anziani et al. 2008). Es erfolgt hierbei keine Notenvergabe, sodass Studierende in einer formativen (im Gegensatz zu einer summativen) Prüfung nicht durchfallen können (Sennhenn-Kirchner et al. 2018). Raupach et al. (2013) untersuchten in einer Studie den Zusammenhang zwischen verschiedenen Prüfungskonsequenzen (summativ vs. formativ) und dem Lernverhalten von Studierenden sowie deren Einfluss auf die Prüfungsleistung der Studierenden. Sie konnten zeigen, dass summative Prüfungen einen starken Einfluss auf das Lernverhalten von Studierenden haben. Dieser Einfluss zeigte sich in Form einer größeren zeitlichen Investition zum Selbststudium sowie der Nutzung von zusätzlichem Lernmaterial. Zudem konnte ein starker Zusammenhang zwischen dem Einsatz von summativen Prüfungen und einer besseren Leistung der Studierenden in einer Abschlussprüfung nachgewiesen werden (Raupach et al. 2013). In einer weiteren Studie konnte gezeigt werden, dass summative Prüfungen die mittelfristige Retention von Fähigkeiten der Studierenden, zur Interpretation eines Elektrokardiogramms (EKG), steigern (Raupach et al. 2016). Hierzu wurde acht Wochen nach der Absolvierung der EKG-Abschlussprüfung ein *retention test* durchgeführt, der zur Überprüfung der zuvor erlernten Fähigkeiten zur EKG-Interpretation dienen sollte. Allerdings zeigte sich, dass der größere Effekt einer summativen (im Vergleich zu einer formativen) Prüfung auf die Fähigkeit zur EKG-Interpretation im Laufe der Zeit abnimmt. Während die Teilnahme an einer summativen Abschlussprüfung einen signfikanten Zusammenhang mit einer besseren Leistung im *retention test* aufwies, war gleichzeitig die Leistungsabnahme (zwischen Abschlussprüfung und *retention test*) von Studierenden, die an den summativen Prüfungen teilnahmen größer als von denjenigen, die die formativen Prüfungen absolvierten. Der mittelfristige im Vergleich zum kurzfristigen Effekt summativer Prüfungen war somit abgeschwächt. Insgesamt zeigten summative Prüfungen einen größeren proportionalen Rückgang der Fähigkeiten zur Interpretation eines EKGs, als formative Prüfungen. Die proportionale Retention der Fähigkeiten zur EKG-Interpretation war damit bei denjenigen Studierenden, die eine summative Prüfung absolvierten, signifikant niedriger (Raupach et al. 2016). Es kann hieraus abgeleitet werden, dass der größere Lernerfolg von summativen im Gegensatz zu formativen Prüfungen häufig nur vorrübergehend ist. Jedoch konnte auch in der zweiten Studie gefolgert werden, dass der Einsatz von summativen Prüfungen einen stärkeren Einfluss auf das Lernverhalten von Studierenden hat, als die Auswahl des Lehrformates (Raupach et al. 2016). Raupach et al.

(2013) schlussfolgerten, dass Lehrende sich des großen Einflusses von (insbesondere summativen) Prüfungen auf das Lernverhalten der Studierenden bewusst sein sollten.

Des Weiteren lässt sich von aggregierten Prüfungsleistungen nur bedingt auf spezifische Stärken und Schwächen innerhalb eines Curriculums schließen (Schiekirka-Schwake et al. 2019). Prüfungsergebnisse stellen somit kein geeignetes Instrument zur Beurteilung der Gesamtheit der Lehrqualität dar und können allenfalls eine Abschätzung des Lehr-Ergebnisses ermöglichen (Schiekirka et al. 2015). Deshalb sollten andere Mittel genutzt werden, die zur Erfassung der Ergebnisqualität geeignet sind. Raupach et al. (2012) schlugen den studentischen Lernerfolg als eine denkbare Messgröße für dieses Ergebnis vor. Diese Form der Ergebnisqualität, im Sinne des Unterschiedes zwischen initialem und abschließendem Leistungsniveau, wird bislang nur in wenigen Evaluationen gemessen (Anders et al. 2016), beispielweise „anhand punktueller studentischer Angaben zum wahrgenommenen eigenen Lernerfolg" (Schiekirka-Schwake et al. 2019). Ein mögliches Konzept zur Bestimmung des studentischen Lernerfolgs wird in Kapitel 1.3.5 ausführlich vorgestellt.

1.3.3 Studentische Selbsteinschätzungen

Studentische Selbsteinschätzungen werden häufig zur Bestimmung der Auswirkungen einer Intervention, beispielweise einer Lehrveranstaltung, genutzt (Drennan und Hyde 2008). Um den Nutzen von studentischen Selbsteinschätzungen genauer betrachten zu können, ist es wichtig, eine Definition von Selbsteinschätzungen zu kennen, die nach Gordon wie folgt beschrieben ist:

> *„Valid self-assessment here means judging ones' performance against appropriate criteria; accurate self-assessment means gaining reasonable concurrence between self-claimed and other, validated, measures of performance"* (Gordon 1991).

Colthart et al. (2008) definierten Selbsteinschätzungen als eine Art „persönliche Evaluation", wobei ein Vergleich der eigenen Fähigkeiten mit einem wahrgenommenen Maßstab erfolgt. Sie dienen sowohl zur Feststellung der eigenen Stärken als auch der Schwächen (Eva und Regehr 2005).

Einige Autoren zeigten, dass die bisherige Forschung Zweifel hinsichtlich der generellen Eignung von Selbsteinschätzungen hegt (Ward et al. 2002; Eva und Regehr 2005). Ward et al. (2002) kritisierten jedoch, dass diese Literatur methodische Probleme aufweist. Des Weiteren definieren die wenigsten Autoren das Konzept der Selbsteinschätzungen, das sie untersuchen (Colthart et al. 2008), sodass ein Vergleich dieser schwierig erscheint. Davis et al. (2006) bemängeln ebenso die Qualität der bisherigen Forschung bezüglich Selbsteinschätzungen unter Medizinern.

Die Fähigkeit, sich selbst in Bezug auf konkrete Fertigkeiten oder Leistungsstände korrekt einzuschätzen ist, im Hinblick auf die spätere ärztliche Tätigkeit, in besonderem Maße wichtig für Medizinstudierende. Metaanalysen konnten zeigen, dass diese Fähigkeit bei

Medizinstudierenden in begrenzter Ausprägung vorhanden ist. In diesen Studien existierte keine signifikante, generelle Über- oder Unterschätzung der Studierenden über alle untersuchten Studien hinweg, jedoch variierten die einzelnen betrachteten Studien diesbezüglich. Zudem steigt die Fähigkeit zur Selbsteinschätzung unter Medizinstudierenden im Laufe des Studiums an (Blanch-Hartigan 2011).

Gordon (1991) untersuchte in einem *Review* 18 verschiedene Studien, die wiederum die Validität von Selbsteinschätzungen behandelten. Es konnte gezeigt werden, dass die Selbsteinschätzungen eigener Leistungen lediglich eine niedrige bis moderate Validität aufweisen und dass diese sich im Laufe des Fortschreitens des Studiums nicht verbessern. Allerdings konnte Gordon (1991) zusammenfassen, dass durchaus Möglichkeiten zur Verbesserung der Validität von Selbsteinschätzungen existieren, indem beispielsweise klare Ziele definiert und Selbsteinschätzungen der Teilnehmer mit den Bewertungen von Außenstehenden verglichen werden.

Es gibt eine Reihe von potenziellen Einflussfaktoren, die zu einer Veränderung der Selbsteinschätzungen führen können. So existiert beispielsweise ein Zusammenhang zwischen den erhobenen Selbsteinschätzungen und der Selbstsicherheit. Eine hohe Selbstsicherheit geht dabei mit einer besseren Selbsteinschätzung einher, jedoch nicht mit einer besseren Leistung (Leopold et al. 2005). Zudem tendieren laut Kruger und Dunning (1999) insbesondere leistungsschwächere Personen, oder solche mit geringen Fähigkeiten in den betrachteten Bereichen dazu, sich selbst zu überschätzen. Die Autoren konnten in vier Studien zeigen, dass Teilnehmer verschiedener Studien die eigenen Leistungen generell höher einschätzten, als die durchschnittliche Leistung aller Teilnehmer. Diese Überschätzung zeigte sich besonders stark bei denjenigen Teilnehmern, deren Leistungen in verschiedenen Tests im unteren Viertel der Kohorte lagen. Die Autoren folgerten, dass denjenigen Personen, die in einem Bereich mangelnde Fähigkeiten aufweisen, zusätzlich die Fähigkeit fehlt, diese Inkompetenz zu erkennen. Aus diesen Gründen muss der Einsatz von Selbsteinschätzungen zur Evaluation gründlich abgewogen werden (Kruger und Dunning 1999).

Braun und Leidner (2009) erklärten, dass Selbsteinschätzungen, die sich auf den Kompetenzerwerb innerhalb eines Kurses beziehen, insgesamt weniger durch *bias* (wie die Beliebtheit der Dozierenden oder die erhaltenen Noten) beeinflusst werden, als Kursbewertungen, welche die Zufriedenheit mit einem Kurs widerspiegeln. Diese Erkenntnis kann ein Anlass dafür sein, den Fokus wie schon in den vorherigen Kapiteln angesprochen, vermehrt auf ergebnisorientierte Evaluationen, auch unter Nutzung von Selbsteinschätzungen, zu legen.

Im Folgenden wird hierzu das Konzept der vergleichenden Selbsteinschätzungen vorgestellt. Dieses soll im Gegensatz zu einzeitig erhobenen Selbsteinschätzungen eine Veränderung messen und dadurch Fortschritt detektieren können. Thompson und Rogers schrieben dazu: *„Self-assessment on expected competencies may be one method for tracking a physician's learning"* (Thompson und Rogers 2008).

Selbsteinschätzungen, die zu einem einzigen Zeitpunkt erhoben werden, bringen teilweise ähnliche Schwierigkeiten mit sich, wie Prüfungen, die nur zu einem Zeitpunkt durchgeführt werden. In beiden Fällen spiegelt das Ergebnis lediglich den aktuellen (selbsteingeschätzten) Wissens- oder Kenntnisstand wider, ohne die Veränderung zu berücksichtigen (Schiekirka et al. 2015). Eine Möglichkeit zur Berücksichtigung und Messung dieser Veränderung ist das Prinzip der „vergleichenden Selbsteinschätzungen". D'Eon et al. (2008) nutzten in einer Studie das Konzept der vergleichenden Selbsteinschätzungen, um den Erfolg eines Workshops zu messen. Diese Studie diente der Validierung des Instrumentes der vergleichenden Selbsteinschätzungen, um die Effektivität einer Intervention, in diesem Fall eines Workshops, zu messen. Der Workshop bezog sich auf die Erteilung von gutem Feedback. Das gemessene Ergebnis am Ende des Workshops war die Veränderung der Fähigkeit, Feedback zu geben. Die Teilnehmenden bewerteten ihre Fähigkeiten durch Selbsteinschätzungen vor (prospektiv) und nach (post) dem Workshop. Zeitgleich zur Selbsteinschätzung nach dem Workshop sollten sie zudem rückblickend ihre Fähigkeiten vor Stattfinden des Workshops einschätzen (retrospektiv). Zusätzlich zu den Selbsteinschätzungen bewerteten zwei unabhängige Beobachter die Fähigkeiten der Teilnehmenden. Die Datenanalyse erfolgte ausschließlich auf Gruppenebene, d. h. durch aggregierte Daten der Teilnehmenden. Sowohl die Selbsteinschätzungen der Teilnehmenden als auch die Bewertungen der externen Beobachter detektierten eine Verbesserung in den Fähigkeiten zur Erteilung von Feedback. Bezogen auf die eigenen Fähigkeiten zur Erteilung von Feedback vor Stattfinden der Intervention, zeigte sich kein signifikanter Unterschied zwischen den prospektiven und retrospektiven Selbsteinschätzungen. Es konnte also kein Einfluss des sog. *response shift bias*, der nachfolgend genauer beschrieben wird, detektiert werden. D'Eon et al. schlussfolgerten, dass vergleichende aggregierte Selbsteinschätzungen eine geeignete Methode sind, um die Effektivität einer Intervention zu messen. Trotz dieser Erkenntnisse raten sie aber davon ab, Entscheidungen allein auf Basis von Selbsteinschätzungen zu treffen (D'Eon et al. 2008). Die Studie wurde jedoch im Anschluss von Lam (2009) kritisiert. Hierbei stellte er zum einen die Generalisierbarkeit der gezogenen Schlüsse in Frage und kritisierte andererseits sowohl die Darstellung als valides Messinstrument trotz erheblicher Über- und Unterschätzungen der Teilnehmenden bezüglich ihrer eigenen Fähigkeiten zur Erteilung von Feedback, als auch die direkte Zurückführung der gemessenen Veränderung auf den Workshop. Er schlussfolgerte, dass Selbsteinschätzungen trotzdem eine Möglichkeit darstellen können, den Erfolg eines Workshops zu messen, wenn bestimmte Voraussetzungen beachtet werden. So sollen u. a. nur valide Selbsteinschätzungen zur Beurteilung des Erfolgs einer Intervention verwendet werden und es muss zudem berücksichtigt werden, dass auch die gleichmäßige Verteilung von Über- und Unterschätzungen zu nicht-validen Selbsteinschätzungen führt (Lam 2009).

Laut Fitzgerald (2003) ist die Fähigkeit von Studierenden zur Selbsteinschätzung über einen langen Zeitraum hinweg stabil. Auch Thompson und Rogers (2008) bestätigten mit ihren Messungen der Selbsteinschätzungen von Studierenden zu mehreren Zeitpunkten den

Nutzen von vergleichenden Selbsteinschätzungen: *„Self-assessment may be a useful method to explore medical students' learning across time"* (Thompson und Rogers 2008).

Nichtsdestotrotz unterliegen auch vergleichende Selbsteinschätzungen verschiedenen Arten von *bias*. Die Kenntnis dieser *bias* ist wichtig, um die Ergebnisse richtig deuten und einordnen zu können. Schiekirka et al. (2014) untersuchten zwei Arten von *bias* genauer, um herauszufinden, ob diese zu einer Beeinflussung der in Kapitel 1.3.5 vorgestellten Lernziel-basierten Evaluationsmethode mittels vergleichender studentischer Selbsteinschätzungen führen können. Im Folgenden werden der sogenannte *response shift bias* und die *implicit theories of change* vorgestellt, um das Verständnis für studentische Selbsteinschätzungen zu vervollständigen und mögliche Schwierigkeiten aufzuzeigen.

Der *response shift bias* wurde ausführlich untersucht und beschrieben (Howard et al. 1979; Howard 1980; Manthei 1997). Dieses Phänomen beschreibt im Hochschulkontext eine Veränderung in der Sichtweise von Studierenden zwischen ihrer initialen Selbsteinschätzung vor (*pretests*) und der Selbsteinschätzung nach einem Kurs oder einer Intervention (*posttests*), die durch eben diese Intervention oder diesen Kurs zustande kommt. Dabei ändert sich das Verständnis des Konstruktes, das sie beurteilen sollen (Drennan und Hyde 2008). Mehrere Autoren empfehlen deshalb retrospektive Einschätzungen (*thentests)* des Leistungsstandes vor dem Kurs, anstelle von wahren Einschätzungen (*pretests*) vor Beginn des Kurses (Howard 1980; Manthei 1997; Skeff et al. 1992). Die sogenannten *thentests* werden dabei, im Gegensatz zu *pretests,* zum gleichen Zeitpunkt erhoben wie die *posttests* nach dem Kurs (Drennan und Hyde 2008). Durch diese zeitliche Nähe werden *thentests* und *posttests* aus der gleichen Perspektive beurteilt und dadurch nicht, wie der Vergleich von *pretests* und *posttests,* durch den *response shift bias* verzerrt (Rohs 1999). Der Einfluss dieses *bias* auf die Evaluationsmethode mittels vergleichender studentischer Selbsteinschätzungen wird in Kapitel 1.3.5 genauer beschrieben.

Implicit theories of change beschreiben hingegen, dass bei der Nutzung von *thentests* und *posttests* Veränderungen in Bewertungen aus dem Grund auftreten können, dass von den Bewertern eine Veränderung erwartet wird, ohne dass diese jedoch tatsächlich stattgefunden haben muss (Norman 2003). Die Schlussfolgerungen von Schiekirka et al. (2014) zum Einfluss dieses *bias* auf die Evaluationsmethode mittels vergleichender studentischer Selbsteinschätzungen werden in Kapitel 1.3.5 zusammengefasst.

1.3.4 Vorhersage-basierte Methode zur Kursevaluation

Wie im vorherigen Kapitel beschrieben, bringen Selbsteinschätzungen von Studierenden mehrere Schwierigkeiten mit sich. Eine generelle Schwierigkeit ist zudem die geringe Teilnahmerate an Evaluationsprozessen (VanGeest et al. 2007). Diese hängt zudem unmittelbar mit der Beeinflussung durch die zuvor genannten *bias* zusammen, denn je höher die Teilnahmerate ist, desto weniger werden die Ergebnisse von *bias* beeinflusst (O'Rourke 1999). Um diesen Schwierigkeiten entgegenzuwirken und dadurch die Aussagekraft der gezogenen

Schlüsse zu erhöhen, haben Forscher aus den Niederlanden einen neuen Ansatz zur Evaluation im Medizinstudium erforscht. Es handelt sich hierbei um die sogenannte Vorhersage-basierte Methode (*prediction-based method*) zur Evaluation eines Kurses (Cohen-Schotanus et al. 2010). Die Autoren adaptierten dazu eine Methode, die ursprünglich aus der Wahlforschung stammt und zur Vorhersage von Wahlergebnissen entwickelt wurde, auf die Evaluation der Lehre eines Kurses an der medizinischen Fakultät Groningen. Hofstee und Schaapmann (1990) führten in ihrer ursprünglichen Studie eine Gegenüberstellung der Ergebnisse professionell durchgeführter Wahlbefragungen mit den Ergebnissen aus geschätzten Vorhersagen von Bürgern zum Ausgang der Wahlen durch. Diese wurden im Anschluss an die Wahlen mit dem tatsächlichen Wahlausgang verglichen. Sie konnten zeigen, dass die Vorhersagen von Laien zum Wahlausgang den professionellen Umfragen, die auf eigenen Meinungen basierten, überlegen sind. Des Weiteren konnte gezeigt werden, dass für die Vorhersagen ein geringerer Rücklauf notwendig ist, um ein reliables Ergebnis zu erhalten, als für die professionellen Wahlbefragungen. Der Vorhersage-basierte Ansatz verfolgt die Idee, dass gemittelte Vorhersagen genauere Ergebnisse liefern als gemittelte eigene Meinungen (Cohen-Schotanus et al. 2010). Cohen-Schotanus et al. (2010) adaptierten diese Methode so auf die Evaluation eines Kurses, dass Studierende anstelle ihrer persönlichen Meinung eines Kurses die Meinung ihrer Kommilitonen über diesen Kurs vorhersagen sollten. Es handelte sich dabei in beiden Fällen um die Erhebung von Globalbewertungen eines Kurses. Zur Berechnung des minimal notwendigen Rücklaufs zum Erhalt reliabler Ergebnisse wurde analog zur Studie von Hofstee und Schaapmann (1990) ein iterativer Vergleich der Ergebnisse einer Subkohorte mit dem Ergebnis der Gesamtkohorte durchgeführt. Diese Vergleiche wurden so häufig in steigender Gruppengröße durchgeführt, bis ein stabiles Ergebnis in Bezug auf die mindestens notwendige Teilnehmerzahl zum Erreichen eines reliablen Ergebnisses, erzielt wurde. Cohen-Schotanus et al. (2010) konnten zeigen, dass die Vorhersage-basierte Methode signifikant weniger Teilnehmer erfordert als die persönlichen Bewertungen der Studierenden, um reliable Ergebnisse zu erhalten. Eine mögliche Erklärung dieser Erkenntnisse ist die Reduzierung persönlicher Faktoren und Auffassungen in der Vorhersage-basierten Methode. Subjektive Faktoren einzelner Studierender werden dabei minimiert. Die Ergebnisse der Vorhersage-basierten Methode repräsentieren somit die Meinung der Gesamtgruppe in geeigneter Weise und eine Anwendung der Vorhersage-basierten Methode im Kontext der Evaluation eines Kurses erscheint sinnvoll (Cohen-Schotanus et al. 2010). Die Studie wurde anschließend von Schönrock-Adema et al. (2013) validiert. Auch in dieser zweiten Studie war zum Erzielen von stabilen, reliablen Ergebnissen eine signifikant niedrigere Rücklaufquote für die Vorhersage-basierte Methode notwendig, als für die Methode, in der die Studierenden ihre persönlichen Globalbewertungen des Kurses abgaben. Eine weitere Schlussfolgerung der Studie war, dass es für die Nutzung der Vorhersage-basierten Methode unwichtig ist, welche Studierenden an der Evaluation teilnehmen, da die Vorhersage-basierte Methode insgesamt weniger von persönlichen Variablen verzerrt wird.

Somit sind bei Verwendung der Vorhersage-basierten Evaluationsmethode keine homogenen Gruppen notwendig (Schönrock-Adema et al. 2013).

Dementsprechend stellt die Vorhersage-basierte Methode zur Evaluation eines Kurses ein geeignetes Instrument für die Messung der studentischen Zufriedenheit mit einem Kurs, im Sinne von Globalbewertungen, dar. Es ist jedoch bisher unklar, ob die Vorhersage-basierte Evaluationsmethode auch zur Messung des studentischen Lernerfolgs geeignet ist.

1.3.5 Evaluation an der UMG

Die Evaluationsordnung der Georg-August-Universität Göttingen schreibt die Evaluation von Lehrveranstaltungen, Studienabschnitten und Studiengängen zur Sicherung der Lehrqualität verbindlich vor (Senat der Georg-August-Universität Göttingen 2006). Aus diesem Grund wurde an der UMG eine Lernziel-basierte Evaluationsmethode entwickelt, bei der mithilfe spezifischer Lernziele der studentische Lernerfolg, im Folgenden auch Lernzuwachs genannt, gemessen werden kann. Die Methode basiert auf vergleichenden Selbsteinschätzungen der Studierenden. Das Lernziel-basierte Evaluationsinstrument mittels vergleichender studentischer Selbsteinschätzungen wurde für alle Module im klinischen Studienabschnitt an der UMG eingeführt (Raupach et al. 2011). Der während eines Moduls erzielte Lernzuwachs ist dabei definiert als Unterschied der Mittelwerte studentischer Selbsteinschätzungen vor (*pre*) und nach (*post*) einem Modul (Raupach et al. 2011). Die Selbsteinschätzungen beziehen sich dabei stets auf spezifische, definierte Lernziele aus dem Göttinger Lernzielkatalog. Indem die Mittelwertdifferenz (*pre* – *post*) durch den korrigierten initialen Mittelwert aller Studierenden dividiert wird, findet mithilfe dieser Methode der initiale Wissens- oder Kenntnisstand der einzelnen Studierenden bei der Berechnung des Lernzuwachses Beachtung (Raupach et al. 2012). Zudem wird durch die zweizeitige Erhebung der tatsächliche Zuwachs, das heißt die stattgefundene Veränderung an Wissen oder Kenntnissen erfasst, die während der Teilnahme am Modul erzielt wurde und nicht nur der aktuelle Stand am Ende eines Moduls. So ergibt sich mithilfe folgender Formel ein prozentualer Lernzuwachs für ein spezifisches Lernziel aus vergleichenden studentischen Selbsteinschätzungen (*comparative self-assessments* = CSA):

$$\text{Lernzuwachs aus CSA [\%]} = \frac{\mu_{pre} - \mu_{post}}{\mu_{pre} - 1} \times 100$$

Hierbei beschreiben μ_{pre} und μ_{post} die Mittelwerte aus studentischen Selbsteinschätzungen für ein spezifisches Lernziel vor (*pre*) bzw. nach (*post*) dem Modul (Raupach et al. 2011).

Diese Methode der Lernerfolgsevaluation wurde mehrfach validiert, sodass die Inhalts- und Kriteriumsvalidität des Instrumentes bestätigt ist (Raupach et al. 2012; Schiekirka et al. 2013; 2014). Die Ergebnisse dieses Evaluationsinstrumentes weisen dabei keine bzw. nur eine schwache Korrelation mit den Ergebnissen traditioneller Evaluationsmethoden, wie etwa den Globalbewertungen eines Kurses, auf. Die Lernerfolgsevaluation spiegelt das Ergebnis, in Form des Lernzuwachses bezogen auf spezifische Lernziele, dabei valide und unabhängig

von Globalbewertungen wider. Somit können traditionelle Evaluationsinstrumente, welche überwiegend strukturelle und prozedurale Aspekte der Lehre bzw. Globalbewertungen eines Kurses erfassen, durch Hinzunahme der Lernerfolgsevaluation um eine neue Dimension in der Evaluation von Lehrveranstaltungen erweitert werden (Raupach et al. 2012). Der Zusammenhang zwischen dem Lernzuwachs aus Selbsteinschätzungen und dem objektiv gemessenen Lernzuwachs ist dabei auf der Gruppenebene stärker als auf der individuellen Ebene einzelner Studierender (Schiekirka et al. 2013).

Die Berechnung des Lernzuwachses anhand der oben genannten Formel beruhte zunächst auf zweizeitigen Datenerhebungen. Die Durchführung von Evaluationen vor und nach jedem Modul ist jedoch zeit- und ressourcenaufwändig. Zudem besteht bei zweizeitigen Datenerhebungen das Risiko, dass die Stichproben sich zu beiden Zeitpunkten nicht gleichen. In einem solchen Fall müsste somit, aufgrund des Risikos der sich unterscheidenden Stichproben, eine paarweise Erhebung der Daten erfolgen. Allerdings wünschen sich Studierende einen geringeren Aufwand bezüglich der Evaluationen (Schiekirka et al. 2012). In einer Folgestudie wurde deshalb der Einsatz von rückblickenden Selbsteinschätzungen am Ende eines Moduls, zur Berechnung des Lernzuwachses, getestet (Schiekirka et al. 2014). Hierbei konnte eine hohe Korrelation zwischen dem prospektiven und dem retrospektiven Lernzuwachs aus Selbsteinschätzungen aufgezeigt werden. Für den prospektiven Lernzuwachs wurden dabei Selbsteinschätzungen zu zwei Zeitpunkten, vor (*pretests*) und nach (*posttests*) dem Modul erhoben. Der retrospektive Lernzuwachs erfordert hingegen nur eine einzeitige Datenerhebung am Ende eines Moduls, da hier der initiale Wissensstand der Studierenden nach Ende des Moduls rückblickend (*thentests*) und der Wissensstand nach dem Modul (*posttest*) an einem Zeitpunkt erhoben werden. Die Selbsteinschätzungen aus *thentests* fielen dabei etwas pessimistischer aus als diejenigen aus den *pretests,* im Sinne des Vorliegens eines *response shift bias* (siehe Kapitel 1.3.3). Die Studierenden bewerteten ihre Leistung rückblickend kritischer und sahen den eigenen Kenntnisstand vor Stattfinden der Lehre geringer an, als sie es vor der Exposition taten. Durch die kritischere Einschätzung ihres initialen Leistungsstandes kommt es insgesamt zu einer Erhöhung des Unterschiedes zwischen initialem und abschließenden Leistungsstand und dadurch zu einem größeren gemessenen Lernzuwachs (Schiekirka et al. 2014). Da dieser Effekt in der beschriebenen Studie in Bezug auf die Berechnung des studentischen Lernzuwachses jedoch nur klein war konnte gefolgert werden, dass beide Methoden gleichwertig sind und somit eine einzeitige Datenerhebung am Ende eines Moduls ausreichend ist, um valide Daten in Hinblick auf studentische Selbsteinschätzungen und den daraus resultierenden Lernzuwachs zu erhalten (Schiekirka et al. 2014). Hierdurch kann der notwendige Aufwand zur Durchführung der Evaluation erheblich reduziert werden.

Schiekirka et al. (2014) untersuchten zudem den Einfluss des *bias implicit theories of change*, der bereits in Kapitel 1.3.3 vorgestellt wurde, in Bezug auf die Evaluationsmethode mittels vergleichender studentischer Selbsteinschätzungen. Sie konnten zeigen, dass der direkte Vergleich rückblickender Selbsteinschätzungen von Studierenden in Form von *thentests* mit den

posttests zu einem leicht größeren Lernzuwachs führte, als wenn kein direkter Vergleich der *thentests* und *posttests* für jedes Lernziel stattfand. Durch den direkten Vergleich kam es zu einem leicht höheren Unterschied zwischen beiden Bewertungen und dadurch zu einem größeren Lernzuwachs. Allerdings zeigte sich auch dieser Effekt in der beschriebenen Studie sehr gering (Schiekirka et al. 2014).

Eine Stärke der beschriebenen Methode der vergleichenden studentischen Selbsteinschätzungen ist die Generierung von Daten zum studentischen Lernerfolg ohne die Notwendigkeit zur Durchführung von Prüfungen vor und nach einem Modul (Anders et al. 2016). Die einzeitige, retrospektive Evaluationsmethode hat sich mittlerweile an der UMG etabliert und wird zur Evaluation am Ende eines jeden Moduls im klinischen Studienabschnitt durchgeführt.

1.4 Ziele dieser Arbeit

In den vorherigen Kapiteln wurde bereits dargelegt, dass valide Evaluationsinstrumente existieren, mithilfe derer eine Lernerfolgsmessung aus wiederholten Selbsteinschätzungen von Studierenden durchgeführt werden kann (Raupach et al. 2011; Raupach et al. 2012; Schiekirka et al. 2013; Schiekirka et al. 2014). Zudem ist gesichert, dass Fremdeinschätzungen von Studierenden, bezogen auf Globalbewertungen eines Kurses, gleichwertig zu den eigenen, globalen Bewertungen der Studierenden über denselben Kurs sind. Dabei erfordern Fremdeinschätzungen, bezogen auf Globalbewertungen, jedoch niedrigere Rücklaufquoten zur Evaluation des Kurses (Cohen-Schotanus et al. 2010; Schönrock-Adema et al. 2013). Offen bleibt bisher, ob diese Erkenntnisse auch auf die Lernerfolgsmessung übertragen werden können. Das Ziel dieser Arbeit war daher die Weiterentwicklung des Instrumentes der lernzielbezogenen Lernerfolgsevaluation mittels vergleichender studentischer Selbsteinschätzungen. Das bisher genutzte Konzept der Lernerfolgsevaluation beruht, wie in Kapitel 1.3.5 beschrieben, auf Selbsteinschätzungen der Studierenden bezogen auf spezifische Lernziele aus dem Göttinger Lernzielkatalog. Die Vorhersage-basierte Evaluationsmethode wurde in Kapitel 1.3.4 vorgestellt. In den Studien basierten die Evaluationen jedoch auf globalen Bewertungen, im Sinne der Zufriedenheit von Studierenden mit dem Kurs. Globale Bewertungen zur Evaluation der Lehre sind jedoch nur begrenzt geeignet, die Gänze der Lehrqualität zu beurteilen. In der vorliegenden Arbeit soll nun der Ansatz der Vorhersage-basierten Evaluationsmethode, im Folgenden auch Fremdeinschätzungs-Methode genannt, auf das Instrument der Lernerfolgsevaluation mittels spezifischer Lernziele übertragen werden. Hiermit soll untersucht werden, ob es einerseits möglich ist, die Fremdeinschätzungs-Methode zur Abschätzung des studentischen Lernerfolgs zu nutzen und andererseits den erforderlichen Mindestrücklauf an Evaluationen zu reduzieren.

Mithilfe der vorliegenden Studie sollen folgende Forschungsfragen beantwortet werden:

1) Wie hängen die mittels der modifizierten Fremdeinschätzungs-Methode ermittelten Lernerfolgsdaten mit den Lernerfolgsdaten der ursprünglich genutzten Selbsteinschätzungs-Methode zusammen?

2) Welcher Rücklauf muss mit beiden Methoden minimal erreicht werden, um reliable Lernerfolgsdaten zu erhalten?

3) Wie hängt der aus Selbsteinschätzungen und Fremdeinschätzungen errechnete Lernerfolg mit dem in wiederholten Prüfungen erhobenen objektiven Lernerfolg zusammen?

2 Material und Methoden

2.1 Studiendesign

Die vorliegende Studie wurde im Wintersemester 2017/2018 im Rahmen des Moduls 3.1 „Erkrankungen des Herz-Kreislauf-Systems und der Lunge“ an der UMG durchgeführt. Dieses Modul ist Teil des dritten klinischen Semesters in Göttingen, welches dem siebten Fachsemester entspricht. Zur Beantwortung der in Kapitel 1.4 genannten Forschungsfragen wurden im Rahmen dieser prospektiven Studie studentische Selbsteinschätzungen, Fremdeinschätzungen sowie Prüfungsleistungen verwendet, um den studentischen Lernerfolg abzuschätzen. Die Studierenden nahmen an zwei unterschiedlichen Zeitpunkten an der Datenerhebung teil (siehe Abbildung 1): Die initiale Datenerhebung fand am ersten Tag des Moduls 3.1 statt (02.10.2017; T1), die finale drei Tage vor Modulende (07.11.2017; T2). Die Prüfung zu T1 war formativer, diejenige zu T2 summativer Natur (s. u.). Diese beiden Prüfungen wurden zusätzlich zur regulären, summativen Modulabschlussklausur (10.11.2017) durchgeführt. Um eine Vergleichbarkeit der objektiven Prüfungen zu erzielen, waren die Inhalte und der Aufbau der Prüfungen an beiden Zeitpunkten identisch. Thematisch unterschieden sich die objektiven Prüfungen dabei grundsätzlich nicht von der regulären Modulabschlussklausur. In beiden Fällen bezogen sich die Prüfungsinhalte auf die Lehrinhalte des Moduls 3.1.

Zur Erhebung der Selbst- und Fremdeinschätzungen wurde den Studierenden jeweils ein zweiteiliger, identischer Fragebogen präsentiert. Zunächst wurden sie gebeten, eine Selbsteinschätzung ihres Leistungsstandes, bezogen auf 29 spezifische Lernziele aus der Projektion des Göttinger Lernzielkatalogs auf das Modul 3.1, durchzuführen. In einem zweiten Schritt wurden die gleichen 29 Lernziele noch einmal präsentiert, jedoch sollten die Studierenden dieses Mal den mittleren Leistungsstand ihrer Kommilitonen in diesen 29 Lernzielen einschätzen. Außerdem wurden demographische Daten wie Alter und Geschlecht sowie die Matrikelnummer erhoben. Die Datensammlung erfolgte über EvaSys® (Electric Paper Evaluationssysteme GmbH, Lüneburg) im Digitalen Prüfungs- und Schulungszentrum der UMG.

Im Anschluss an die Erhebung der Selbst- und Fremdeinschätzungen wurde an beiden Terminen (T1 und T2) die objektive Prüfung durchgeführt. Diese bezog sich auf die gleichen 29 Lernziele wie die Aussagen zur Selbst- und Fremdeinschätzung.

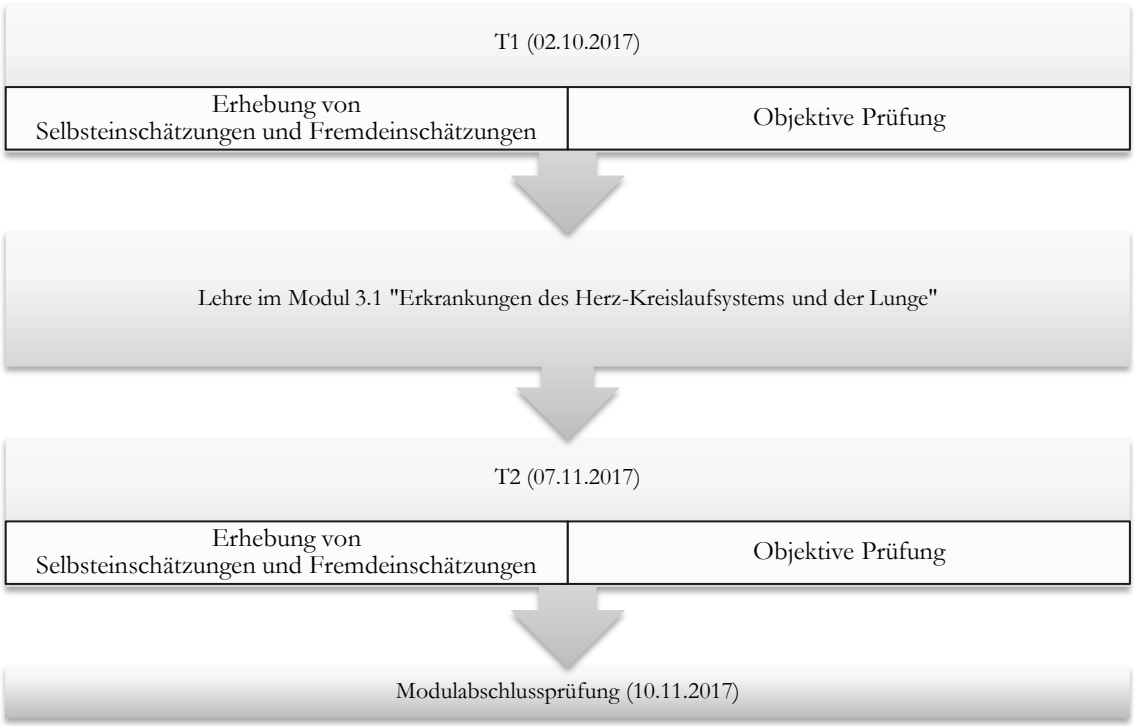

Abbildung 1: Aufbau der Studie. Dargestellt ist der zeitliche Ablauf des Forschungsprojektes.

2.1.1 Probandenrekrutierung

Zur Studienteilnahme wurden alle Studierenden eingeladen, die sich im Wintersemester 2017/2018 im dritten klinischen Semester befanden und sich zur Teilnahme am Modul 3.1 angemeldet hatten. Die Studierenden wurden im Vorfeld per E-Mail über die Möglichkeit zur Studienteilnahme und die damit einhergehende Nutzung ihrer Daten zu Studienzwecken informiert. Für die Termine T1 und T2 bestand eine Anwesenheitspflicht der Studierenden, da die Datenerhebung unabhängig vom begleitenden Forschungsprojekt vor allem der Planung und der Qualitätssicherung der modularen Lehre diente. An diesen Terminen wurde die Studie erneut vorgestellt und Informationsblätter für Probanden an die Studierenden verteilt. Wesentliches Einschlusskriterium für die Studie war die vollständige Teilnahme an den Datenerhebungen. Zudem musste zum Einschluss in die Studie eine schriftliche Einverständniserklärung zur Studienteilnahme vorliegen.

2.1.2 Leistungsanreize

Durch die Teilnahme an der objektiven Prüfung zum Zeitpunkt T2 konnten bis zu fünf Leistungspunkte für das Fach 11 (F11) „Innere Medizin“ erzielt werden. Insgesamt können im Modul 3.1 25 Leistungspunkte für F11 erworben werden; dies entspricht 25 % aller für dieses Fach im klinischen Studienabschnitt erreichbaren Punkte. Der Anreiz, bis zu fünf Leistungspunkte zu erwerben, entsprach also einem Anteil von 20 % der zu erwerbenden Leistungspunkte in diesem Modul und einem Anteil von 5 % aller Leistungspunkte für das F11 insgesamt (entsprechend einer halben Notenstufe). Dieses Vorgehen sollte die

Motivation der Studierenden zur ernsthaften Bearbeitung der Prüfungsfragen, erhöhen. Zusätzlich wurden zehn Büchergutscheine im Wert von je 25 Euro für die zehn Personen als Belohnung ausgesetzt, deren Fremdeinschätzungen zum Zeitpunkt T2 am genauesten mit dem wahren Ergebnis der summativen Prüfung des Abschlusstestates der Gesamtkohorte übereinstimmten. Dieser Anreiz sollte zu einer möglichst genauen Einschätzung des Leistungsstandes der Kommilitonen führen.

Die objektive Prüfung bestand aus 145 *True-False-Items*, dementsprechend konnten bis zu 145 Rohpunkte erzielt werden. Aufgrund der hohen Ratewahrscheinlichkeit in dieser Prüfungsform wurde im Vorhinein festgelegt, dass Leistungspunkte erst ab einer Erfolgsquote von mindestens 50 % vergeben wurden. Die weitere Staffelung der Leistungspunkte erfolgte in Schritten von 10 %. Studierende, die nicht an der Studie teilnehmen wollten, konnten die Leistungspunkte unabhängig davon im Rahmen der Prüfung zum Zeitpunkt T2 erwerben.

2.2 Durchführung

2.2.1 Selbsteinschätzungen

Um die Selbsteinschätzungen der Studierenden zu deren aktuellem Wissensstand zu ermitteln, wurde für jedes der 29 spezifischen Lernziele eine Selbsteinschätzungsaussage formuliert. Die Studierenden wurden gebeten, diese Aussagen einzeln zu bewerten. Dabei konnten sie auf einer sechsstufigen Likert-Skala ihre aktuellen Fähigkeiten oder ihren aktuellen Wissensstand zu dem spezifischen Lernziel (zum Zeitpunkt T1 bzw. T2) von „trifft voll zu“ bis „trifft überhaupt nicht zu“ einschätzen (siehe Abbildung 2). Es wurden hieraus Mittelwerte über die gesamte Kohorte gebildet, um den aggregierten Lernzuwachs auf Gruppenebene zu berechnen, den die Studierenden durch die Selbsteinschätzungen angegeben haben. Dies war zur Beantwortung der Forschungsfragen eins und drei notwendig. Zusätzlich wurde auf Grundlage der Selbsteinschätzungen für jedes Lernziel der individuelle Lernzuwachs jedes einzelnen Studierenden berechnet (Formeln siehe Kapitel 2.3.1).

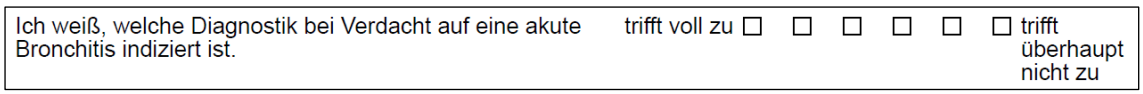

Abbildung 2: Aussage zur Selbsteinschätzung für das Lernziel „Akute Bronchitis“. Aussage für die Selbsteinschätzung zu einem spezifischen Lernziel auf einer sechsstufigen Likert-Skala. Die Studierenden konnten Angaben von „trifft voll zu“ bis „trifft überhaupt nicht zu“ machen. Diese Angaben wurden anschließend in Zahlenwerte von Eins (entspricht der Aussage: „trifft voll zu“) bis Sechs (entspricht der Aussage: „trifft überhaupt nicht zu“) umkodiert.

2.2.2 Fremdeinschätzungen

Analog zu den Selbsteinschätzungen wurden die oben bereits erwähnten Fremdeinschätzungen erhoben. Die hierbei genutzten Aussagen bezogen sich auf die gleichen 29 Lernziele. Die Aussagen, die die Studierenden bewerten sollten, waren dabei identisch mit denen der Selbsteinschätzungen, jedoch sollte hierbei der mittlere Fähigkeiten- oder Wissensstand der Kommilitonen bewertet werden. Es wurde dafür erneut für jede Aussage eine sechsstufige Likert-Skala erstellt, auf der die Studierenden ihre Einschätzung von „trifft voll zu“ bis „trifft überhaupt nicht zu“ einordnen konnten (siehe Abbildung 3). Zusätzlich zu den geschilderten Berechnungen aus Kapitel 2.2.1 wurden für die Beantwortung der Forschungsfragen eins und drei die Mittelwerte der Punktzahlen aus den Fremdeinschätzungen aller Studierenden berechnet. Diese dienten der Berechnung des auf Fremdeinschätzungen basierenden aggregierten Lernzuwachses. Des Weiteren wurde für jedes Lernziel ein individueller Lernzuwachs berechnet, der sich aus den Fremdeinschätzungen jedes einzelnden Studierenden in Bezug auf die Kommilitonen ergab (Formeln siehe Kapitel 2.3.1).

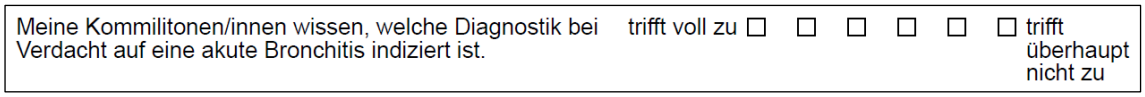

Abbildung 3: Aussage zur Fremdeinschätzung für das Lernziel „Akute Bronchitis“. Aussage für die Fremdeinschätzung zu einem spezifischen Lernziel auf einer sechsstufigen Likert-Skala. Die Studierenden konnten Angaben von „trifft voll zu“ bis „trifft überhaupt nicht zu“ machen. Diese Angaben wurden anschließend in Zahlenwerte von Eins (entspricht der Aussage: „trifft voll zu“) bis Sechs (entspricht der Aussage: „trifft überhaupt nicht zu“) umkodiert.

2.2.3 Objektive Prüfungsleistungen

Um zu vergleichen, ob der Lernzuwachs der Studierenden basierend auf Selbst- und Fremdeinschätzungen eine Korrelation zum tatsächlichen Lernzuwachs während des Moduls aufweist, war es notwendig, diesen Lernzuwachs objektiv zu bestimmen. Hierzu wurde an beiden Terminen (T1 und T2) eine objektive Prüfung durchgeführt. Diese war inhaltlich an beiden Zeitpunkten identisch. Sie bestand aus 29 Fragen, die auf den gleichen 29 Lernzielen basierten wie die Selbst- und Fremdeinschätzungsaussagen. Um die Kongruenz zwischen Lehre und Prüfung zu gewährleisten, wurde der Inhalt des Moduls von der Autorin vorab aufgearbeitet und entsprechende Prüfungsfragen formuliert. Diese Fragen wurden vor der endgültigen Auswahl durch den Projektleiter validiert und eine Musterlösung erstellt. 28 der 29 Fragen bestanden dabei jeweils aus fünf *True-False-Items,* im Folgenden als Unteritems bezeichnet, die die Studierenden einzeln bewerten sollten. Für jedes einzelne Unteritem sollte dabei entschieden werden, ob die jeweilige Aussage „richtig“ oder „falsch“ ist (siehe Abbildung 4). Auf diese Weise konnten die Studierenden pro Lernziel zwischen null und fünf Punkte erreichen. Bei einer Frage sollten die Studierenden in einem Freitextfeld bis zu fünf Antworten auf die Frage eintragen, welche Organe/Organsysteme in erster Linie von

Folgeerkrankungen der arteriellen Hypertonie betroffen sein können (siehe Abbildung 5). Somit konnten auch bei dieser Aufgabenstellung zwischen null und fünf Punkten erreicht werden. Um eine Vergleichbarkeit von Selbst- und Fremdeinschätzungen sowie der objektiven Prüfung zu ermöglichen, musste zunächst eine Anpassung der Skalen erfolgen. Hierzu wurden die Ergebnisse der objektiven Prüfung invertiert und in Werte auf einer Skala von Eins bis Sechs übertragen. Die sechsstufige Einteilung ermöglichte die Umrechnung der Punkte auf die ebenfalls sechsstufige Likert-Skala des Erhebungsinstruments für die Selbst- und Fremdeinschätzungen. Ein Wert von fünf Punkten in der objektiven Prüfung entsprach somit nach der Anpassung einem Wert von Eins, ein Wert von null Punkten in der Prüfung wurde in einen Wert von Sechs umkodiert. Somit spiegelte das Erreichen der vollen Punktzahl von fünf Punkten in der objektiven Prüfung die Skalenoption „trifft voll zu“ in den Selbst- und Fremdeinschätzungen.

Im Folgenden sind einige diagnostische Maßnahmen aufgelistet. Bitte kennzeichnen Sie durch Ankreuzen für jede dieser Maßnahmen, ob sie jeweils bei einem ansonsten gesunden Patienten mit V.a. eine akute Bronchitis indiziert ist („Richtig“) oder nicht („Falsch“).
(Bitte entscheiden Sie bei **jeder** Aussage, ob diese zutrifft oder nicht!)

(A) Ja ☒ Kleines Blutbild
(B) Ja ☒ CRP-Bestimmung
(C) Ja ☒ Lungenfunktion
(D) Ja ☒ Sputumdiagnostik
(E) ☒ Nein Lungenauskultation

Abbildung 4: Frage aus der objektiven Prüfung zum Lernziel „Akute Bronchitis“ mit *True-False-Items* und Musterlösung. Die Studierenden sollten für jedes einzelne Unteritem (A – E) entscheiden, ob dieses im Hinblick auf die im Einleitungstext genannte Aussage „richtig“ oder „falsch“ ist. Jedes korrekt beantwortete Unteritem ergab einen Rohpunkt. Es konnten somit pro Frage und Lernziel zwischen null und fünf Rohpunkten erreicht werden. Die korrekten Antworten sind in der Abbildung durch ein Kreuz markiert.

Bitte tragen Sie in das folgende Feld die fünf Organe/Organsysteme ein, die in erster Linie von Folgeerkrankungen der arteriellen Hypertonie betroffen sein können.

✓ Herz
✓ (Ge)Hirn
✓ Niere(n)
✓ Auge(n)
✓ Gefäße

Abbildung 5: Freitextfrage aus der objektiven Prüfung zum Lernziel „Hypertonie“ mit Musterlösung. Die Studierenden sollten bei dieser Aufgabe bis zu fünf Antworten in das Freitextfeld eintragen. Jede korrekte Antwort ergab einen Rohpunkt, sodass auch bei dieser Aufgabenstellung bis zu fünf Rohpunkte erreicht werden konnten. Falsche Antworten führten nicht zu einem Punktabzug. Zusätzlich zur Aufgabenstellung sind in der Abbildung die Musterlösungen für die fünf Antworten dargestellt.

Aus den invertierten Werten der objektiven Prüfungen wurden Mittelwerte über die Gesamtkohorte gebildet, um den objektiven aggregierten Lernzuwachs zu berechnen (Formeln siehe Kapitel 2.3.1). Zudem wurde für jeden Studierenden der objektive individuelle Lernzuwachs bestimmt.

2.3 Datenanalyse

2.3.1 Lernzuwachsraten

Um den Lernzuwachs berechnen und damit die Forschungsfragen eins und drei beantworten zu können, wurde die in Göttingen entwickelte Formel zur Berechnung des studentischen Lernerfolges mithilfe vergleichender Selbsteinschätzungen genutzt (Raupach et al. 2011). Zur Bestimmung des Lernzuwachses wurden die Werte aus Selbsteinschätzungen (siehe Kapitel 2.2.1), Fremdeinschätzungen (siehe Kapitel 2.2.2) und der objektiven Prüfung (siehe Kapitel 2.2.3) genutzt.

Einerseits kann dabei der Lernzuwachs auf Gruppenebene betrachtet werden. Hierbei werden Mittelwerte der Gesamtkohorte gebildet und in die weiter unten aufgeführte Formel für den aggregierten Lernzuwachs (*aggregated performance gain* = APG) eingesetzt. Diese Formel kann sowohl für den Lernzuwachs in der objektiven Prüfung als auch für den Lernzuwachs, den die Studierenden laut Selbst- oder Fremdeinschätzungen erreicht haben, genutzt werden. Beim Lernzuwachs, der sich aus den objektiven Prüfungen ergab, handelt es sich um den objektiven Lernzuwachs (*objective aggregated performance gain* = oAPG). Der Lernzuwachs, der laut Selbsteinschätzungen erzielt wurde, wird im Folgenden als subjektiver Lernzuwachs (*subjective aggregated performance gain* = sAPG) bezeichnet. Der Lernzuwachs basierend auf den Fremdeinschätzungen ist als vorhersage-basierter Lernzuwachs (*prediction-based aggregated performance gain* = pAPG) gekennzeichnet.

Die zu T1 erhobenen Mittelwerte der objektiven Prüfung, Selbst- oder Fremdeinschätzungen sind im Folgenden durch $\bar{x}_{pre}$ gekennzeichnet. Die Mittelwerte zu T2 werden in der Formel durch $\bar{x}_{post}$ gekennzeichnet. Die Formel für den aggregierten Lernerfolg (APG) lautet:

$$\text{APG } [\%] = \frac{\bar{x}_{pre} - \bar{x}_{post}}{\bar{x}_{pre} - 1} \times 100$$

(Quelle: Schiekirka et al. (2013))

Nach Einsetzen der Werte in die Formel ergaben sich dadurch pro Lernziel drei Werte, die den Lernzuwachsraten entsprachen und in Prozent angegeben wurden. Bei Verwendung der Mittelwerte der objektiven Prüfung ergab sich daraus der oAPG-Wert, bei Verwendung der Mittelwerte der Selbst- oder Fremdeinschätzungen entsprechend der sAPG- bzw. pAPG-Wert.

Wie in den vorherigen Kapiteln beschrieben, wurde zusätzlich zur Analyse des Lernzuwachses auf Gruppenebene der individuelle Lernzuwachs (*individual performance gain* = IPG) für jeden einzelnen Studierenden für ein spezifisches Lernziel berechnet. Auch hier wurde jeweils pro Lernziel ein Lernzuwachs aus der objektiven Prüfung (*objective individual performance gain* = oIPG), laut Selbsteinschätzung (*subjective individual performance gain* = sIPG) sowie laut Fremdeinschätzung (*predicition-based individual performance gain* = pIPG) berechnet. Durch einen paarweisen Vergleich der *pre*- und *post*-Werte eines einzelnen Studierenden kann der IPG berechnet werden. x_{pre} steht dabei für den initialen Wert vor Beginn des Moduls (T1), x_{post} für den Wert nach dem Modul (T2). Die Formel zur Berechnung des individuellen Lernzuwachses lautet:

$$\text{IPG } [\%] = \frac{x_{pre} - x_{post}}{x_{pre} - 1} \times 100$$

(Quelle: Schiekirka et al. (2013))

2.3.2 Vergleich von Selbsteinschätzungs-Methode und Fremdeinschätzungs-Methode

Die erste Forschungsfrage bezog sich darauf, ob durch die Selbsteinschätzungs- und die Fremdeinschätzungs-Methode vergleichbare Daten zum Lernzuwachs generiert werden. Hierzu wurde die Pearson-Korrelation zwischen sAPG und pAPG berechnet. Nach Cohen (1988) entspricht $|r| = 0{,}1$ einem kleinen Effekt, $|r| = 0{,}3$ einem mittleren Effekt und $|r| = 0{,}5$ einem großen Effekt.

Zur Beantwortung der ersten Forschungsfrage wurde zudem ein Bland-Altman-Plot erstellt, der ein Instrument zum Vergleich zweier Messmethoden darstellt (Bland und Altman 1999). Bland und Altman (1986) beschrieben, dass eine hohe Korrelation zwischen zwei Messmethoden nicht automatisch einer guten Übereinstimmung dieser beiden Methoden entspricht. Sie erklärten dazu, dass der Korrelationskoeffizient r lediglich die Stärke des Zusammenhangs zweier Messmethoden, jedoch nicht deren exakte Übereinstimmung misst. Eine exakte Übereinstimmung zwischen zwei Methoden würde nur in einem solchen Fall vorliegen, in dem die Werte sich entlang einer Geraden aufreihen würden, die den Nullpunkt beider Achsen schneidet (entsprechend einem Intercept von Null). Eine sehr hohe bzw. perfekte Korrelation würde sich hingegen in jedem Fall ergeben, in dem die Variablen sich entlang einer beliebigen Geraden (mit einem beliebigen Intercept) aufreihen würden (Bland und Altman 1986). Deshalb entwickelten sie einen Ansatz zum Vergleich zweier Messmethoden der Auskunft darüber liefert, inwiefern sich eine neue Methode von einer altbewährten unterscheidet (Bland und Altman 1999). Der Grundgedanke dabei ist, die Unterschiede zwischen einzelnen Messwerten zu quantifizieren. Da jede Art von Messung Fehler mit sich bringt, ist es in vielen Fällen nicht möglich zu bestimmen, welcher der beiden Tests oder Messmethoden den wahren Wert genauer misst. Deshalb legten Altman und Bland (1983) den Fokus auf die Darstellung der Vergleichbarkeit zweier Methoden. In einem Bland-

Altman-Plot wird für jedes Item die Differenz der Messwerte beider Methoden auf der Ordinate gegen den Mittelwert der Messwerte beider Methoden auf der Abszisse aufgetragen. Durch diese Darstellung können sowohl die Höhe des Unterschiedes zwischen beiden Methoden abgeschätzt werden als auch Ausreißer identifiziert werden. Zudem können Trends, in Form von größeren oder kleineren Differenzen zwischen den Methoden bei hohen oder niedrigen Werten, leichter erkannt werden. Die Tendenz einer Methode, generell höher oder niedriger zu messen als die andere Methode, kann anhand der mittleren Differenz beider Methoden, auch *bias* genannt, abgeschätzt werden. Im Idealfall würde dieser *bias,* der graphisch durch eine horizontale Linie im Diagramm markiert wird, nahe dem Wert Null liegen. Altman und Bland (1983) führten diesen graphischen Vergleich für Messmethoden ein, die im medizinischen Kontext eine Quantität messen (Altman und Bland 1983; Bland und Altman 1999).

In der vorliegenden Arbeit wurde zum Vergleich der Fremdeinschätzungs- und der Selbsteinschätzungs-Methode ein Bland-Altman-Plot erstellt. Dabei wurde die Differenz des prozentualen Lernzuwachses von beiden Methoden (entspricht in dieser Studie: sAPG – pAPG) für jedes einzelne Lernziel gegen den Mittelwert aus beiden Methoden derselben Lernziele (entspricht in diesem Fall: (sAPG + pAPG)/2) dargestellt. Zur Beurteilung wurden zudem drei Hilfslinien eingezeichnet. Die Werte für diese Hilfslinien entsprechen einerseits der mittleren prozentualen Differenz aller Lernziele, andererseits der Standardabweichung der Mittelwertdifferenz multipliziert mit dem Faktor 1,96, wobei dieser Wert einmal zur Differenz der Mittelwerte addiert, das andere Mal davon subtrahiert wird. Man bezeichnet den Bereich zwischen diesen Linien als 95 % *limits of agreement* (Bland und Altman 1999). Idealerweise streuen die Messwerte in einem Bland-Altman-Plot nahe um die horizontale Hilfslinie, die die mittlere Differenz zwischen den Messwerten der beiden Methoden repräsentiert. Im besten Fall würde sich dabei keine Tendenz in höheren oder niedrigeren Messwerten abzeichnen. Es ist jedoch auch möglich, dass sich die beiden Methoden systematisch voneinander unterscheiden, je höher oder niedriger die Messwerte ausfallen. In diesem Fall würde es sich um einen „proportionalen“ *bias* handeln, bei dem der Unterschied zwischen den Messwerten abhängig von der Höhe der Messwerte sinkt oder steigt (Bland und Altman 1999). Im Bland-Altman-Plot würden sich die Messwerte dann entlang einer Funktion aufreihen, die von der Horizontalen abweicht. Abweichungen von der mittleren Differenz bei besonders hohen oder niedrigen Messwerten würden darauf hinweisen, dass die beiden Messmethoden bei Extremwerten stark voneinander differieren und hier eine Methode erheblich höher oder niedriger misst als die andere. Die Vergleichbarkeit der beiden Methoden wäre somit bei Extremwerten eingeschränkt.

2.3.3 Vergleich der Selbsteinschätzungs- und Fremdeinschätzungs-Methode mit den objektiven Prüfungsleistungen

Die dritte Forschungsfrage beschäftigt sich mit dem Vergleich des Lernzuwachses der Selbsteinschätzungs- und der Fremdeinschätzungs-Methode mit dem Lernzuwachs aus der objektiven Prüfung. Hierzu wurden Korrelationskoeffizienten zwischen dem Lernzuwachs aus Selbsteinschätzungen und objektiven Prüfungsleistungen sowie Fremdeinschätzungen und objektiven Prüfungsleistungen berechnet. Des Weiteren wurden Bland-Altman-Plots erstellt, um die Vergleichbarkeit von Fremdeinschätzungs-Methode und objektiver Prüfung sowie Selbsteinschätzungs-Methode und objektiver Prüfung darzustellen.

2.3.4 Berechnung des erforderlichen Mindestrücklaufs

Die zweite Forschungsfrage bezog sich darauf, welcher Mindestrücklauf an Selbst- oder Fremdeinschätzungen erforderlich ist, um reliable Daten über den Lernzuwachs zu erhalten. Dazu wurde der Ansatz von Hofstee & Schaapmann (1990) verwendet. Bei dieser Methode erfolgt der wiederholte Vergleich des Ergebnisses einer Subkohorte mit dem Ergebnis der Gesamtkohorte. Wiederholt bedeutet in diesem Fall, dass (begonnen mit einer Subkohorte bestehend aus einem Probanden) bei jeder wiederholten Messung ein weiterer Proband in die vorherige Subkohorte einbezogen wird und in jedem Schritt das Ergebnis dieser neu gebildeten Subkohorte mit dem Ergebnis der Gesamtkohorte verglichen wird. Die Auswahl der Probanden, die in die Subkohorte einbezogen werden, erfolgt zufällig. Durch diese Vorgehensweise steigt mit jedem weiteren Vergleich die Größe der zufällig gebildeten Subkohorte. So soll die minimale Anzahl an Probanden identifiziert werden die notwendig ist, um ein Ergebnis zu erhalten, das dem der Gesamtkohorte weitgehend gleicht. In der vorliegenden Studie wurde wiederholt der prozentuale Lernzuwachs aus Selbst- und Fremdeinschätzungen einer immer größer werdenden Subkohorte mit dem prozentualen Lernzuwachs der Gesamtkohorte verglichen. So sollte für beide Methoden für jedes einzelne Lernziel die Anzahl an Studierenden identifiziert werden, die mindestens notwendig ist, um reliable Daten zu erhalten (*number needed to evaluate* = NNE). Hierbei wurde definiert, dass ab dem Punkt ein stabiles Ergebnis vorliegt, an dem die Abweichung des Ergebnisses der Subkohorte vom Ergebnis der Gesamtkohorte (entspricht der Einbeziehung von 100 % aller Probanden) absolut < 5 % betrug. Die Analysen wurden sowohl für die Selbst- als auch für die Fremdeinschätzungs-Methode durchgeführt, sodass die Ergebnisse anschließend verglichen werden konnten, um zu überprüfen, ob für beide Methoden der gleiche Rücklauf notwendig ist. Des Weiteren wurde für beide Methoden graphisch die NNE für jedes einzelne Lernziel mit dem dazugehörigen oAPG aufgetragen. Hierdurch konnte untersucht werden, ob die NNE variiert, wenn der objektiv gemessene Lernzuwachs besonders hoch oder niedrig ausfällt.

2.3.5 Statistische Auswertung

Die statistischen Analysen wurden mithilfe des Statistikprogramms IBM SPSS Statistics für Windows, Version 26 (IBM Corp., Armonk, New York, USA) durchgeführt. Ergänzend wurde das Programm Microsoft Excel 2016 für Windows, Version 16.0 verwendet. Zudem wurde zur Berechnung des erforderlichen Mindestrücklaufs sowie zur Durchführung der ROC-Analyse das Statistikprogramm R (R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org) genutzt.

Zur Überprüfung einer Normalverteilung der Mittelwerte aus Selbsteinschätzungen, Fremdeinschätzungen und objektiven Prüfungsleistungen wurde der Kolmogorov-Smirnov-Test (ohne Lilliefors-Korrektur) verwendet.

Zur deskriptiven Analyse der Daten wurden, wenn nicht anders beschrieben, Mittelwerte (MW) und Standardabweichungen bestimmt. Für die MW der Punktzahlen aus Selbsteinschätzungen, Fremdeinschätzungen und objektiven Prüfungsleistungen wurden Standardfehler bestimmt. Das Signifikanzlevel wurde auf 5 % festgesetzt. Zudem wurde für die objektiven Prüfungen eine Itemanalyse durchgeführt. Diese wurde auf der Ebene der Unteritems berechnet. Hierzu wurden Itemschwierigkeiten und Itemtrennschärfen berechnet und die interne Konsistenz mittels Cronbach's α bestimmt (Möltner et al. 2006).

Weiterhin wurde überprüft, ob signifikante Veränderungen der Selbsteinschätzungen, Fremdeinschätzungen und objektiven Prüfungsleistungen zwischen T1 und T2 stattgefunden haben. Bei Vorliegen einer Normalverteilung wurden gepaarte t-Tests verwendet, andernfalls wurde der Wilcoxon-Test genutzt. Diese Tests wurden auf die MW der Selbsteinschätzungen, Fremdeinschätzungen und objektiven Prüfungsleistungen zum Zeitpunkt T1 und T2 angewendet. Darüber hinaus wurde eine Bonferroni-Korrektur durchgeführt. Diese Analysen dienten damit dem Vergleich beider Methoden (Forschungsfrage eins) sowie zum Vergleich von sAPG und pAPG mit den objektiven Prüfungsleistungen (Forschungsfrage drei).

Zum Vergleich der minimal notwendigen Rücklaufquoten der Selbst- und Fremdeinschätzungs-Methode wurden die NNE für beide Methoden auf das Vorliegen einer Normalverteilung überprüft und in Abhängigkeit davon ein t-Test durchgeführt.

In diesem Zusammenhang soll für das Verständnis der weiteren Analysen der Begriff „diagnostischer Test" erläutert werden. Ein diagnostischer Test in der klinischen Medizin entscheidet, ob ein Individuum als „gesund" oder „krank" klassifiziert wird (Armitage et al. 2002). Um zu beurteilen, inwiefern ein Test in der Lage ist diese Klassifikation korrekt durchzuführen, gibt es verschiedene Kenngrößen. Die Sensitivität beschreibt die „Richtig-positiv"-Rate, das heißt Individuen, die krank sind, werden auch als solche durch den Test erkannt. Die Spezifität entspricht der „Richtig-negativ"-Rate und umfasst diejenigen Individuen, die vom Test als „gesund" klassifiziert werden und auch tatsächlich gesund sind (Honest und

Khan 2002). Der positiv prädiktive Wert (PPW) entspricht dem Anteil der tatsächlich Kranken unter allen als „krank" klassifizierten Individuen, während der negativ prädiktive Wert (NPW) dem Anteil der tatsächlich Gesunden unter allen denjenigen beschreibt, die vom Test als „gesund" klassifiziert wurden (Armitage et al. 2002). Dabei existieren diagnostische Entscheidungen auch außerhalb der klinischen Medizin. In diesen Fällen wird mithilfe eines diagnostischen Tests die Unterscheidung zwischen „ja" oder „nein", anstelle von „gesund" oder „krank" vorgenommen (Swets et al. 2000).

Zur Bestimmung von Grenzwerten zur Unterscheidung zwischen Lernzielen mit optimalem und suboptimalem Lernzuwachs (Kriteriumsvalidität) wurde eine *Receiver-operating-characteristics*-Analyse (ROC-Analyse) durchgeführt. Eine ROC-Analyse dient der Bestimmung von Grenzwerten, im Folgenden *Cut-off*-Werte genannt, in diagnostischen Entscheidungen (Swets et al. 2000). Jede Art von diagnostischen Entscheidungen erfordert die Festlegung von solchen *Cut-off*-Werten, mithilfe derer in der Medizin beispielsweise zwischen „krank" und „nicht krank" unterschieden wird, in anderen Bereichen zwischen „ja" und „nein". Eine ROC-Analyse ist eine mathematische Herangehensweise an die Festlegung dieser *Cut-off*-Werte, um falsch-positive und falsch-negative Entscheidungen so gering wie möglich zu halten (Swets et al. 2000). Dabei zeigt die Kurve die verschiedenen Szenarien auf, die sich aus unterschiedlich gewählten *Cut-off*-Werten ergeben. Diese Szenarien werden durch die richtig-positiv Rate (entspricht der Sensitivität) sowie die falsch-positiv Rate (entspricht 1 – Spezifität) widergespiegelt (Swets et al. 2000). Die ROC-Analyse stellt dazu in einem Koordinatensystem unterschiedliche Szenarien als Wertepaar der Sensitivität (auf der Ordinate) und der Spezifität (als 1 – Spezifität auf der Abszisse) eines Tests dar. Jeder Punkt repräsentiert dabei einen möglichen *Cut-off*-Wert und die sich daraus ergebenden Werte für die Sensitivität und Spezifität des Tests. Die Werte auf der Ordinate entsprechen damit dem richtig-positiven Anteil, die Werte auf der Abszisse dem falsch-positiven Anteil (Zweig und Campbell 1993). Für diagnostische Tests werden prinzipiell eine hohe Sensitivität (entsprechend hohen Werten auf der Ordinate) und eine hohe Spezifität (entsprechend niedrigen Werten auf der Abszisse) angestrebt. Je weiter die aus den Wertepaaren gebildete Kurve sich im Diagramm der linken und oberen Grenze annähert, desto besser ist der Test zur Differenzierung geeignet (Metz 1978). In der vorliegenden Studie wurde sowohl für die Selbst- als auch für die Fremdeinschätzungs-Methode eine ROC-Analyse durchgeführt. Neben den *Cut-off*-Werten wurde die *area under the curve* (AUC) für beide Methoden bestimmt. Die Sensitivität entsprach dabei den optimal gelehrten Lernzielen, die durch die jeweilige Methode als „optimal gelehrt" identifiziert wurden. „Optimal" bedeutete in diesem Zusammenhang, dass der aggregierte Lernzuwachs in der objektiven Prüfung mindestens 50 % betrug. Die Spezifität entsprach den „suboptimal erreichten Lernzielen", die durch die jeweilige Methode als solche erkannt wurden.

2.3.6 Sensitivitätsanalyse

Im Anschluss an die primäre Datenanalyse wurde ausgehend von den Lernzuwachs-Ergebnissen zu T2 eine Sensitivitätsanalyse durchgeführt, um die inhaltliche Passung der Items zu überprüfen. Die Sensitivitätsanalyse war nicht von vorneherein als Auswertungsschritt vorgesehen. Da sich nach Durchführung der Sensitivitätsanalyse der für die weiteren Auswertungen genutzte Datensatz änderte, bildet das Kapitel zur Sensitivitätsanalyse deshalb den Anfang des Ergebnisteils dieser Arbeit (siehe Kapitel 3.2). Die Analyse wurde für alle Items durchgeführt, bei denen eine Differenz von mehr als 20 % zwischen sAPG oder pAPG einerseits und oAPG andererseits bestand. Dazu wurden von der Autorin, dem Studienleiter und einer weiteren Mitarbeiterin des Bereichs Medizindidaktik und Ausbildungsforschung der UMG die Selbst- und Fremdeinschätzungsaussagen sowie die Prüfungsfragen auf Kongruenz überprüft. Des Weiteren wurden bei Auffälligkeiten alle zu den Themen der betreffenden Items zur Verfügung stehenden Lehr- und Lernmaterialien überprüft. Dies beinhaltete die Analyse von Podcasts und Vorlesungsfolien, die den Studierenden zum eigenständigen Lernen zur Verfügung standen. Hierdurch sollte ebenfalls eine Kongruenz zwischen gelehrten und geprüften Inhalten sichergestellt werden. Bei fehlender Kongruenz, d. h. in der Formulierung voneinander abweichenden Selbsteinschätzungs-/Fremdeinschätzungsaussagen und Prüfungsitems, wurde entschieden, die Frage aus der Wertung zu nehmen. Grund dafür war die Feststellung, dass die Studierenden andere Inhalte evaluierten, als durch die objektive Prüfung getestet wurden. Bei fehlender Kongruenz zwischen gelehrten Inhalten und den in der objektiven Prüfung als korrekt gewerteten Antworten, wurden diese Items umkodiert. Durch diese Umkodierung wurde ein *True-False-Item* als korrekt beantwortet gewertet, wenn es von den Studierenden so beurteilt wurde, wie es in der Vorlesung gelehrt wurde (obwohl es bei der Prüfungskonzeption anders intendiert war). Bei fehlenden Hinweisen auf eine inhaltliche oder anderweitig fehlende Kongruenz und keiner ausreichenden Erklärung für die Abweichung einer Frage wurde diese unverändert in der Wertung belassen.

2.4 Aufklärung und Ethikvotum

Im Rahmen der Studie wurden personenbezogene Daten erhoben. Zum Zusammenführen der Daten beider Zeitpunkte (T1 und T2) sowie zur korrekten Zuordnung der erworbenen Leistungspunkte aus der Prüfung zu T2, war eine Erhebung zusammen mit einer eindeutigen Kennung (hier: Matrikelnummer) notwendig. Nach dem Zusammenführen der Datensätze und der Vergabe der Leistungspunkte sowie der Gutscheine wurde diese gelöscht. So lag vor Beginn der Auswertung ein anonymisierter Datensatz vor. Das Studienprotokoll wurde der Ethikkommission der Universitätsmedizin Göttingen vorgelegt (Antragsnummer 14/9/16 mit Votum vom 21.09.2016).

3 Ergebnisse

Im Folgenden werden zunächst die Charakteristika der teilnehmenden Studierenden beschrieben. Anschließend werden die Ergebnisse der Analysen zur Hypothesentestung bezüglich der Forschungsfragen dargestellt.

3.1 Charakterisierung der Stichprobe

Insgesamt waren im Wintersemester 2017/2018 146 Studierende für das Modul 3.1 angemeldet, wovon 130 Studierende in die Studienteilnahme einwilligten. Zwei Studierende nahmen nur an einem der beiden Zeitpunkte (T1 oder T2) an der Datenerhebung teil. Diese Daten konnten aufgrund der Unvollständigkeit nicht in die Analyse einfließen. Es ergaben sich 128 vollständige Datensätze, entsprechend einer Rücklaufquote von 87,7 %. Der Anteil der weiblichen Studierenden betrug dabei 64,1 % (82 Personen) und der Anteil der männlichen Studierenden 35,9 % (46 Personen). Die Alterspanne betrug am ersten Tag des Moduls zwischen 20 und 39 Jahren, mit einem Mittelwert von 23,9 ± 3,1 Jahren.

3.2 Sensitivitätsanalyse

Bei fünf der 29 Items bestand zum Zeitpunkt T2 eine Differenz zwischen dem prozentualen Lernzuwachs aus Selbst- bzw. Fremdeinschätzungen und dem aus der objektiven Prüfung von mehr als 20 % (siehe Tabelle 1). Bei den übrigen 24 Items bestand im Mittel eine Differenz von 2,9 ± 10,1 % zwischen dem Lernzuwachs aus Fremdeinschätzungen und objektiver Prüfung sowie von 2,9 ± 11,0 % zwischen dem Lernzuwachs aus Selbsteinschätzungen und objektiver Prüfung. Es wurde daraufhin nur für die fünf abweichenden Items eine Sensitivitätsanalyse durchgeführt. Das Ergebnis der Analyse ergab, dass zwei der Items aus der Wertung genommen wurden, zwei Items umkodiert wurden und ein Item unverändert in der Auswertung belassen wurde. Es ergab sich somit nach der Sensitivitätsanalyse ein neuer, umkodierter Datensatz mit nun 27 Items. Im Folgenden sind die Ergebnisse der Sensitivitätsanalyse der fünf abweichenden Items detailliert dargestellt. Die Ergebnisse der Itemanalyse der fünf Items und deren Unteritems vor der Umkodierung sind im Anhang detailliert aufgeführt (siehe Anhang A1 – A4).

Tabelle 1: Lernzuwachsraten der fünf abweichenden Items

Item	Lernziel	Lernzuwachs		
		Objektive Prüfung	Selbsteinschätzung	Fremdeinschätzung
7	Perikarditis	31,1 %	76,8 %	77,9 %
24	Koronare Herzkrankheit	39,2 %	81,1 %	83,2 %
25	Kardiomyopathie	10,1 %	68,9 %	71,7 %
28	Herzinsuffizienz	55,1 %	84,2 %	76,6 %
29	Fallot'sche Tetralogie	35,4 %	70,2 %	73,6 %

Dargestellt sind die fünf abweichenden Items mit dem dazugehörigen Lernzuwachs der Selbsteinschätzungen, Fremdeinschätzungen und objektiven Prüfung sowie das Lernziel, auf das sie sich beziehen.

Bezogen auf das Lernziel „Perikarditis" wurde im Rahmen der Sensitivitätsanalyse folgendes Problem identifiziert: Zu drei der fünf Unteritems gab es Hinweise in der durchgeführten Lehre, die Anlass dazu gaben, bezüglich der richtigen Antwort in die eine bzw. andere Richtung zu argumentieren. Da im Nachhinein nicht nachvollzogen werden konnte, was in den Vorlesungen gesagt bzw. als „korrekt" gelehrt wurde, wurde diese Frage aus der Wertung genommen. Die Studierenden konnten aufgrund uneindeutiger Lehrinhalte möglicherweise nicht klar zwischen den Antwortmöglichkeiten differenzieren.

Für das Lernziel „Koronare Herzkrankheit" zeigte sich, dass keine ausreichende Kongruenz zwischen der Selbst- und Fremdeinschätzungsaussage und der zugehörigen Frage aus der objektiven Prüfung bestand. Es fiel der Entschluss, dieses Item ebenfalls aus der Wertung zu nehmen.

In Bezug auf das Lernziel „Kardiomyopathie" ergab sich weder ein Hinweis auf Inkongruenzen zwischen Lehre und Prüfung noch auf Fehler in der Item-Konstruktion. Dementsprechend wurde das Item unverändert in der Analyse belassen.

Für das Lernziel „Herzinsuffizienz" konnte ein Unteritem identifiziert werden, das lediglich von 11,7 % der Studierenden richtig beantwortet wurde, während die anderen vier Unteritems von > 95 % richtig beantwortet wurden. Bei diesem abweichenden Unteritem konnte eine Diskrepanz zwischen der als „richtig" gewerteten Antwortmöglichkeit in der objektiven Prüfung und der Lehre im Modul (aufgrund von Podcasts und Vorlesungsfolien, die den Studierenden zur Verfügung standen) identifiziert werden. Somit wurde entschieden, das abweichende Unteritem umzukodieren und die Antwort als „richtig" zu werten, die der Lehre entsprach. Die Umkodierung bedeutete, dass die zuvor als „richtig" gewertete Antwortmöglichkeit des Unteritems nach der Umkodierung als „falsch" gewertet wurde und umgekehrt. Der Lernzuwachs in der objektiven Prüfung betrug somit vor der Umkodierung des Unteritems 55,1 % und im Anschluss 88,6 %.

Es zeigte sich auch für das Lernziel „Fallot'sche Tetralogie", dass vier der Unteritems von jeweils > 70 % der Studierenden richtig beantwortet wurden, während das erste Unteritem lediglich von 11,7 % der Studierenden korrekt beantwortet wurde. Die Sensitivitätsanalyse ergab, dass auch hier eine inhaltliche Diskrepanz zwischen durchgeführter Lehre und der als „richtig" gewerteten Antwortmöglichkeit bestand. Somit fiel der Entschluss zur Umkodierung des Unteritems. Diese erfolgte auf die gleiche Weise wie beim Lernziel „Herzinsuffizienz" beschrieben. Nach der Umkodierung entsprach der Lernzuwachs in der objektiven Prüfung 61,9 %, im Vergleich zu 35,4 % vor der Umkodierung.

3.3 Vergleich von Selbsteinschätzungs-Methode und Fremdeinschätzungs-Methode (Forschungsfrage eins)

Um die Forschungsfrage eins (Wie hängen die mittels der modifizierten Fremdeinschätzungs-Methode ermittelten Lernerfolgsdaten mit den Lernerfolgsdaten der ursprünglich genutzten Selbsteinschätzungs-Methode zusammen?) zu beantworten, wurde der Lernzuwachs aus den aggregierten Selbst- und Fremdeinschätzungen berechnet. Die MW der Punktzahlen für die Selbst- und Fremdeinschätzungen vor und nach dem Modul sind in Abbildung 6 und Abbildung 7 dargestellt. Hier lässt sich erkennen, dass die MW der Selbsteinschätzungen vor dem Modul eine große Spannweite aufwiesen (2,19 – 5,72). Den höchsten numerischen Wert wies das Lernziel „Fallot'sche Tetralogie" auf. Im Hinblick auf dieses Lernziel haben die Studierenden somit vor dem Modul laut Selbsteinschätzungen das geringste Vorwissen gehabt. Den niedrigsten Wert wies das Lernziel „Atherosklerose" auf, dementsprechend hatten die Studierenden ihrer eigenen Ansicht nach vor Beginn des Moduls in Bezug auf dieses Lernziel bereits das größte Vorwissen. Nach dem Ende des Moduls war die Spannweite der mittleren Punktzahlen geringer und betrug zwischen 1,31 und 2,56. In Bezug auf die Fremdeinschätzungen lag eine ähnliche Situation vor. Auch hier war die Spannweite der Mittelwerte vor dem Modul groß (1,93 – 5,32). In diesem Fall schätzten die Studierenden das Vorwissen ihrer Kommilitonen in Bezug auf das Lernziel „Pneumonie" am geringsten ein. Für das Lernziel „Atherosklerose" hingegen schätzten sie, dass ihre Kommilitonen bereits das größte Vorwissen besäßen. Wie bereits für die Mittelwerte der Selbsteinschätzungen beschrieben, war auch die Spannweite der mittleren Punktzahlen der Fremdeinschätzungen nach dem Modul geringer (1,26 – 2,23).

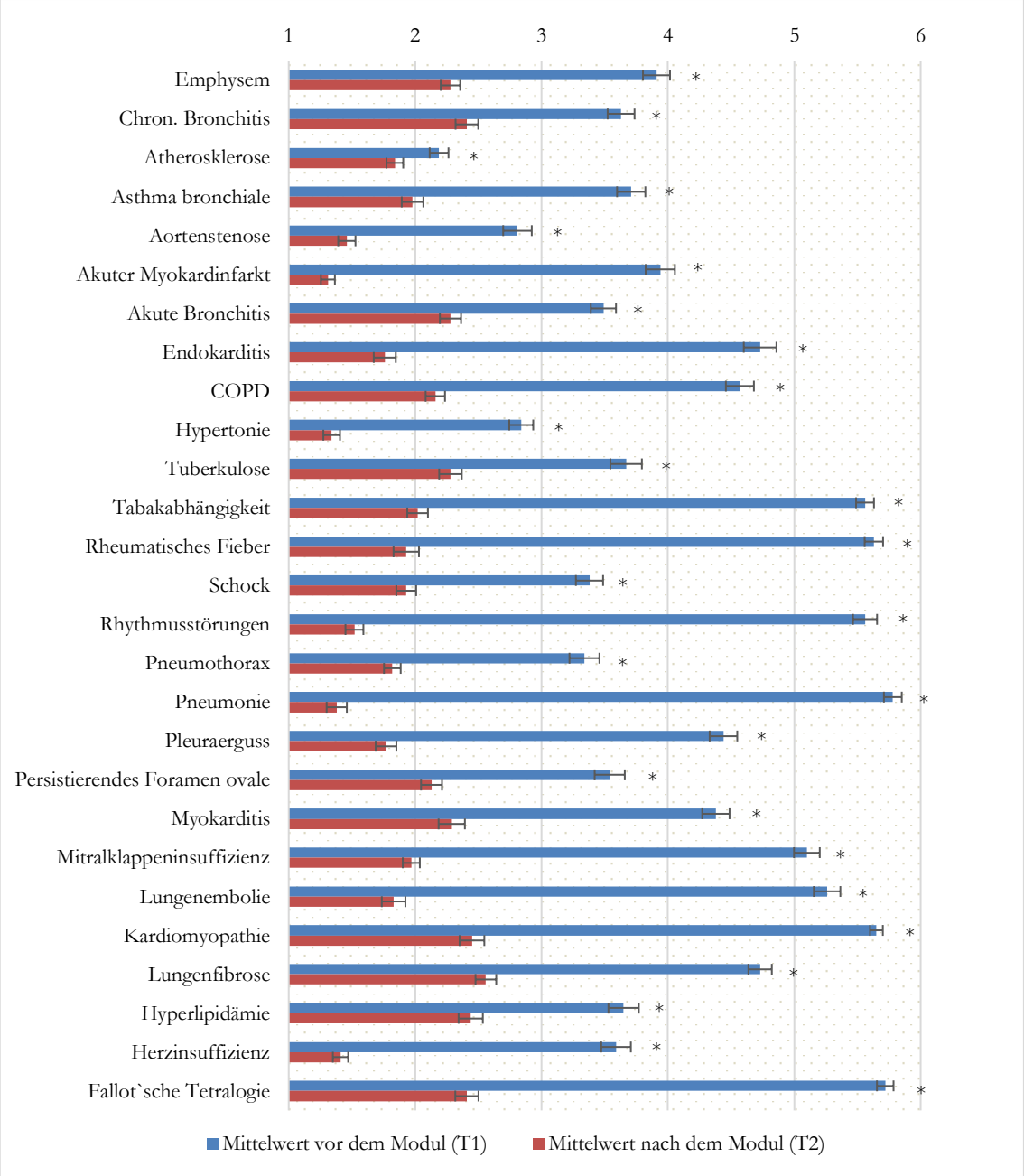

Abbildung 6: Selbsteinschätzungen für jedes Lernziel vor (T1) und nach (T2) dem Modul. Abgebildet sind die Mittelwerte sowie Standardfehler der Punktzahlen pro Lernziel, die die Studierenden als Selbsteinschätzungen auf einer sechsstufigen Likert-Skala angegeben haben. Die Mittelwerte zum Zeitpunkt T1 sind in Form von blauen, diejenigen zum Zeitpunkt T2 in Form von roten Balken dargestellt. Es konnten Angaben von „trifft voll zu“ (entspricht einem Zahlenwert von Eins) bis „trifft überhaupt nicht zu“ (entspricht einem Zahlenwert von Sechs) gemacht werden. Zudem ist jedes Lernziel, das nach Bonferroni-Korrektur einen signifikanten Unterschied der MW zwischen T1 und T2 im Wilcoxon-Test zeigte, mit einem „*“ markiert.

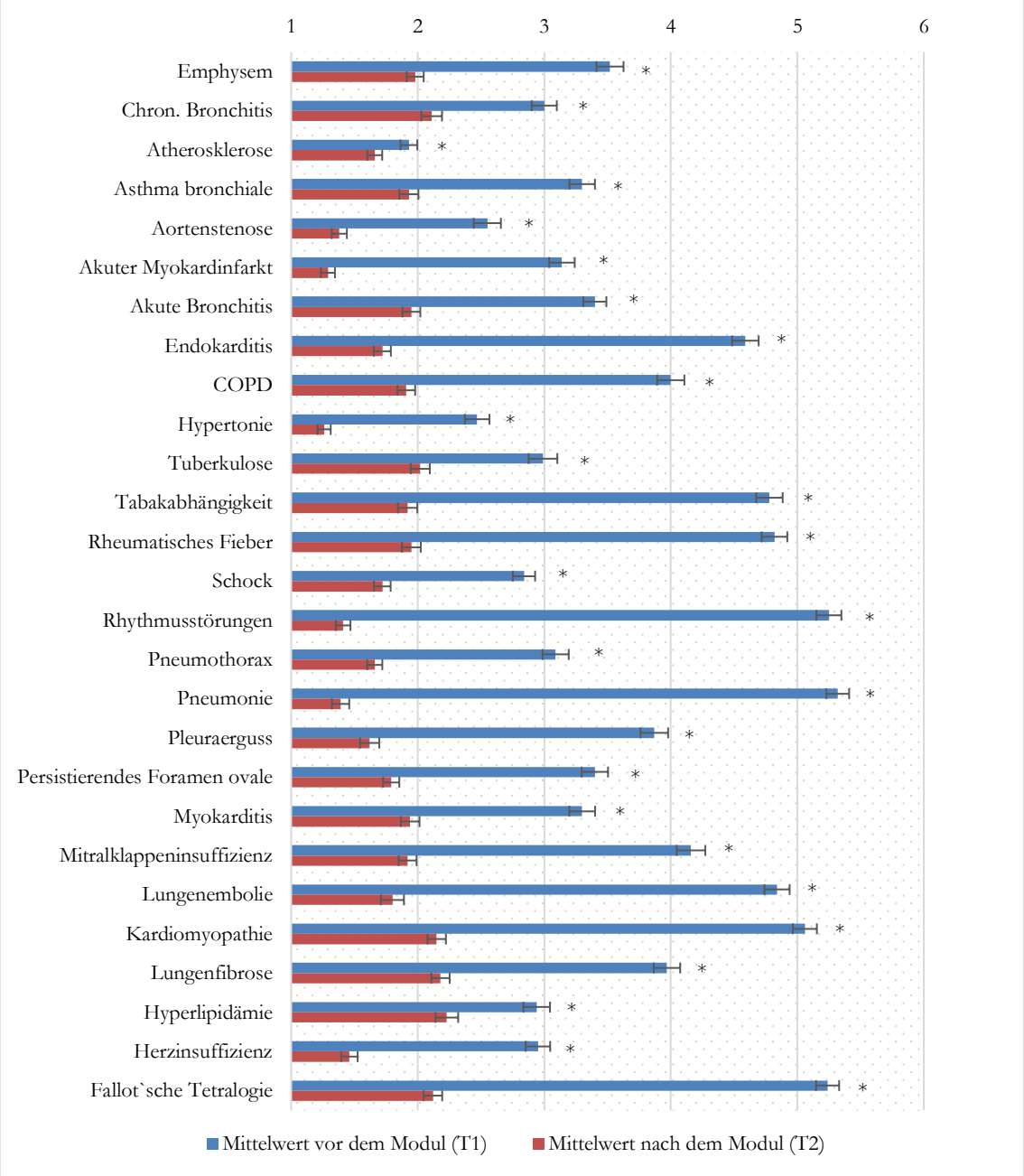

Abbildung 7: Fremdeinschätzungen für jedes Lernziel vor (T1) und nach (T2) dem Modul. Abgebildet sind die Mittelwerte sowie Standardfehler der Punktzahlen pro Lernziel, die die Studierenden als Fremdeinschätzungen auf einer sechsstufigen Likert-Skala angegeben haben. Die Mittelwerte zum Zeitpunkt T1 sind in Form von blauen, diejenigen zum Zeitpunkt T2 in Form von roten Balken dargestellt. Es konnten Angaben von „trifft voll zu“ (entspricht einem Zahlenwert von Eins) bis „trifft überhaupt nicht zu“ (entspricht einem Zahlenwert von Sechs) gemacht werden. Zudem ist jedes Lernziel, das nach Bonferroni-Korrektur einen signifikanten Unterschied der MW zwischen T1 und T2 im Wilcoxon-Test zeigte, mit einem „*“ markiert.

Die Annahme einer Normalverteilung wurde für die Selbsteinschätzungen und Fremdeinschätzungen getestet. Diese konnte in beiden Fällen für keines der 27 Lernziele bestätigt werden, sodass daraufhin der Wilcoxon-Test für zwei verbundene Stichproben genutzt wurde. Nach Durchführung der Bonferroni-Korrektur ergab sich ein korrigierter p-Wert von

0,00185. Es zeigte sich für die Selbsteinschätzungen für jedes der 27 Lernziele ein signifikanter Unterschied zwischen T1 und T2. Auch die Fremdeinschätzungen für die 27 Lernziele unterschieden sich zu T1 und T2 signifikant.

Nach Einsetzen der Mittelwerte in die Formel für den APG entstand für jedes Lernziel ein mittlerer prozentualer Lernzuwachs, der sich aus den Selbsteinschätzungen ergab und ein mittlerer prozentualer Lernzuwachs, der sich aus den Fremdeinschätzungen ergab. Die so berechneten Werte für den Lernzuwachs wurden genutzt, um die Forschungsfrage eins zu beantworten. Zudem werden sie im Kapitel 3.5 erneut aufgegriffen, um Forschungsfrage drei zu bearbeiten. Beispielhaft soll hier einmal die Berechnung des aggregierten Lernzuwachses laut Selbsteinschätzungen und Fremdeinschätzungen für das Lernziel „Akute Bronchitis" dargestellt werden:

Ermittelter Lernzuwachs basierend auf den Selbsteinschätzungen für das Lernziel „Akute Bronchitis":

$$\text{sAPG [\%]} = \frac{3{,}49 - 2{,}28}{3{,}49 - 1} \times 100 = 48{,}6$$

Ermittelter Lernzuwachs basierend auf den Fremdeinschätzungen für das Lernziel „Akute Bronchitis":

$$\text{pAPG [\%]} = \frac{3{,}4 - 1{,}95}{3{,}4 - 1} \times 100 = 60{,}4$$

Für das Lernziel „Akute Bronchitis" betrugt der ermittelte Lernzuwachs laut Selbsteinschätzungen demnach 48,6 %, laut Fremdeinschätzungen 60,4 %.

Bezogen auf alle 27 Lernziele zeigte sich zwischen dem sAPG und dem pAPG eine starke Korrelation mit $r = 0{,}952$ ($p < 0{,}001$), entsprechend einer Varianzaufklärung von knapp 91 % (siehe Abbildung 8).

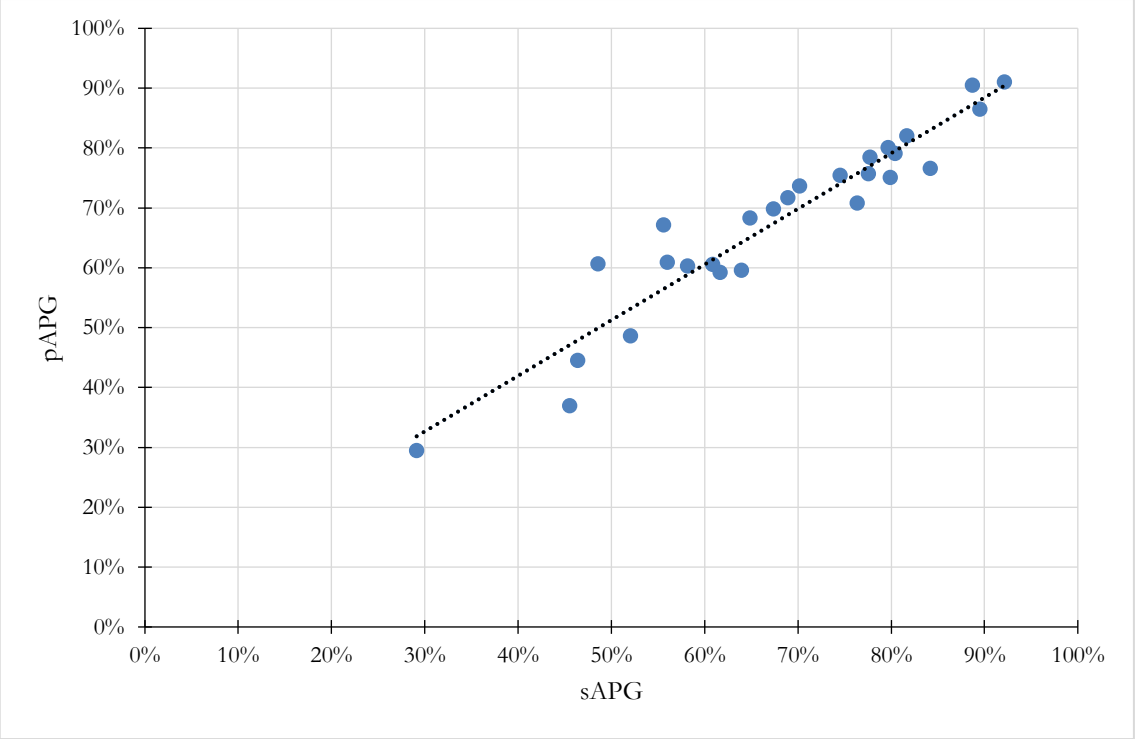

Abbildung 8: Korrelation des Lernzuwachses aus Selbst- und Fremdeinschätzungen. Dargestellt ist der Lernzuwachs in Prozent basierend auf Selbsteinschätzungen und Fremdeinschätzungen. Jeder einzelne Datenpunkt entspricht dabei einem der 27 Lernziele.

Der mittlere aggregierte Lernzuwachs aus allen 27 Items betrug für die Selbsteinschätzungs-Methode 67,8 % (Minimum: 29,1 %; Maximum: 92,2 %) und für die Fremdeinschätzungs-Methode ebenfalls 67,8 % (Minimum: 29,4 %; Maximum: 91,0 %). Die Standardabweichung der Selbsteinschätzungen betrug 15,6 %. Bei den Fremdeinschätzungen betrug die Standardabweichung 15,2 %.

Zum Vergleich der beiden Methoden wurde zudem ein Bland-Altman-Plot (siehe Abbildung 9) erstellt. Die mittlere Differenz des Lernzuwachses pro Lernziel zwischen den beiden Methoden betrug 0,002 %. Abgesehen von zwei Items zu den Lernzielen „Akute Bronchitis“ und „Persistierendes Foramen ovale“ lagen alle anderen innerhalb des 95 % *limits of agreement*.

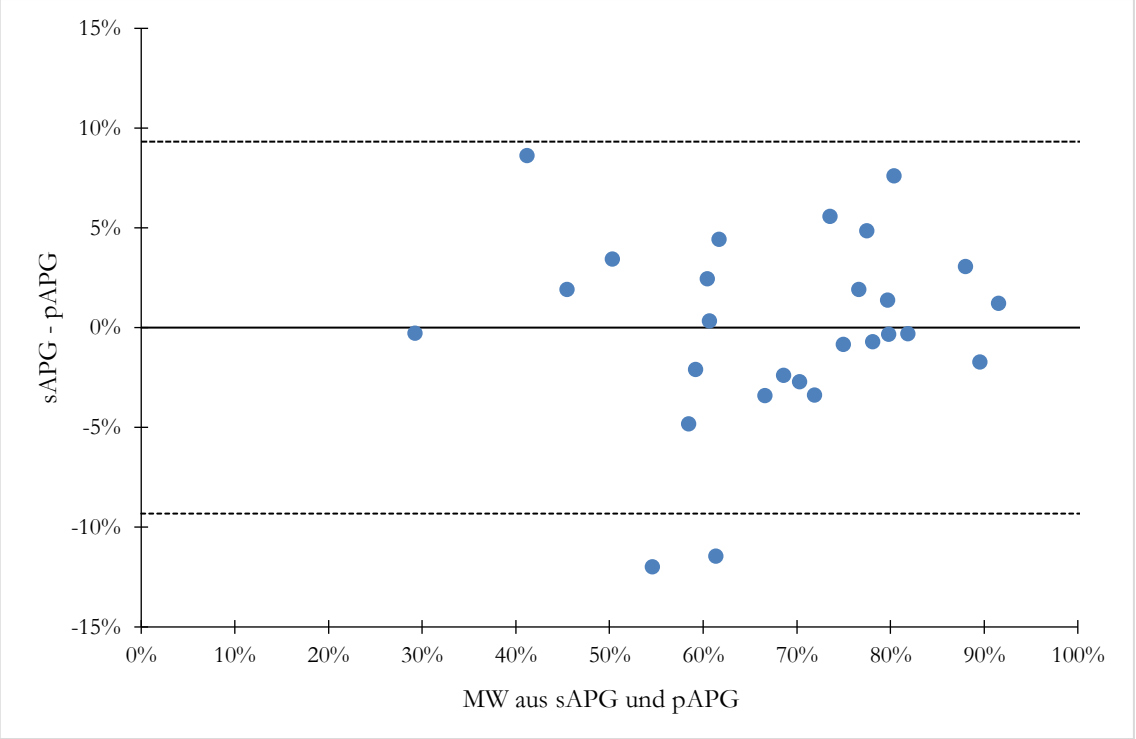

Abbildung 9: Bland-Altman-Plot des Lernzuwachses aus Selbst- und Fremdeinschätzungen. Jeder Datenpunkt entspricht einem Lernziel. Zudem sind drei waagerechte Hilfslinien eingezeichnet. Die durchgezogene Linie entspricht dem MW der Differenzen von sAPG und pAPG aller 27 Lernziele in Prozent. Die gestrichelten Linien entsprechen diesem MW ± 1,96 × Standardabweichung des MW der Differenz.

Auf Grundlage der hier gezeigten Daten kann gefolgert werden, dass die Selbsteinschätzungs-Methode und die Fremdeinschätzungs-Methode vergleichbare Lernerfolgsdaten liefern.

3.4 Erforderlicher Rücklauf für Selbsteinschätzungs- und Fremdeinschätzungs-Methode (Forschungsfrage zwei)

Im folgenden Abschnitt soll Forschungsfrage zwei (Welcher Rücklauf muss mit beiden Methoden minimal erreicht werden, um reliable Lernerfolgsdaten zu erhalten?) beantwortet werden. Wie in Kapitel 2.3.4 beschrieben, wurde der Ansatz von Hofstee & Schaapmann (1990) verwendet und iterative Vergleiche des mittleren Ergebnisses der Selbst- bzw. Fremdeinschätzungen einer zufällig gebildeten Subkohorte mit dem mittleren Ergebnis der Gesamtkohorte durchgeführt. Diese Vergleiche wurden für jedes einzelne Lernziel durchgeführt. Abbildung 10 zeigt beispielhaft diesen Prozess für das Lernziel „Akute Bronchitis“.

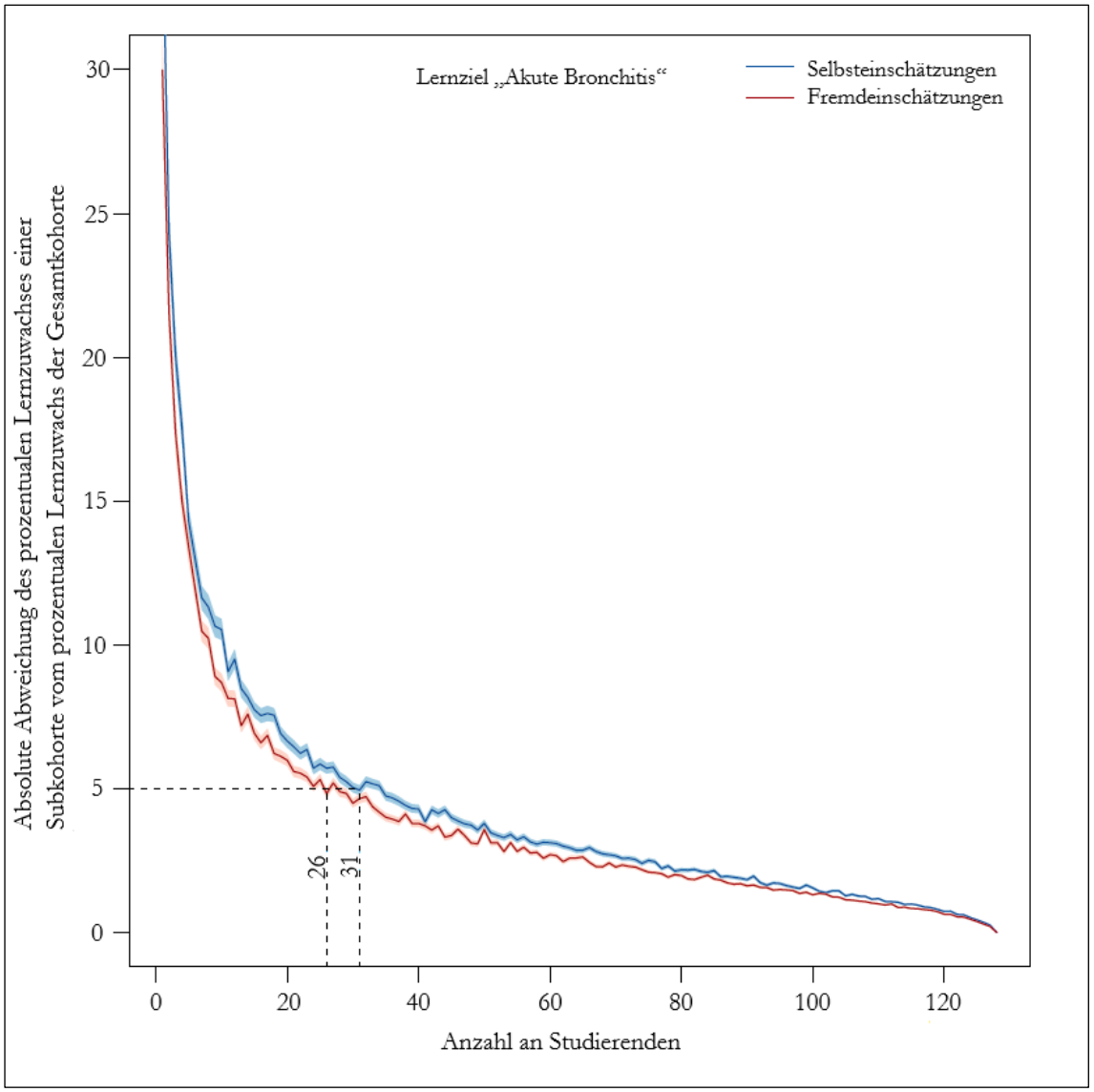

Abbildung 10: Iterativer Vergleich des prozentualen Lernzuwachses einer Subkohorte mit dem prozentualen Lernzuwachs der Gesamtkohorte für das Lernziel „Akute Bronchitis“. Für dieses Lernziel sind mindestens 26 Studierende notwendig, um ein reliables Ergebnis für den auf Fremdeinschätzungen basierenden Lernzuwachs zu erhalten. Für Selbsteinschätzungen beträgt diese Anzahl an Studierenden 31. Der Grenzwert wurde auf < 5 % Abweichung zwischen dem Ergebnis der Subkohorte und dem der Gesamtkohorte festgesetzt.

Tabelle 2 zeigt die NNE für jedes Lernziel für die Selbst- und Fremdeinschätzungen, die notwendig ist, um ein reliables Ergebnis zu erzielen, das mit dem auf Selbsteinschätzungen bzw. Fremdeinschätzungen basierenden Lernzuwachs der Gesamtkohorte vergleichbar ist.

Tabelle 2: NNE für Selbst- und Fremdeinschätzungen

Lernziel	Selbsteinschätzungen	Fremdeinschätzungen
Emphysem	24	24
Chronische Bronchitis	35	47
Atherosklerose	64	69
Asthma bronchiale	29	28
Aortenstenose	30	34
Akuter Myokardinfarkt	9	15
Akute Bronchitis	31	26
Endokarditis	17	14
COPD	15	19
Hypertonie	34	31
Tuberkulose	32	45
Tabakabhängigkeit	11	14
Rheumatisches Fieber	16	12
Schock	32	39
Rhythmusstörungen	9	6
Pneumothorax	24	26
Pneumonie	12	8
Pleuraerguss	18	22
Persistierendes Foramen ovale	35	24
Myokarditis	29	32
Mitralklappeninsuffizienz	10	19
Lungenembolie	14	17
Kardiomyopathie	35	11
Lungenfibrose	16	20
Hyperlipidämie	42	51
Herzinsuffizienz	14	26
Fallot‘sche Tetralogie	19	9

Dargestellt ist die NNE für jedes einzelne Lernziel für die Selbst- und die Fremdeinschätzungs-Methode.

Die im Mittel notwendige Zahl an Studierenden, um ein reliables Ergebnis zu erhalten, betrug für die Selbsteinschätzungen 24,3 ± 3,5 Studierende (Minimum: 9; Maximum: 64). Für die Fremdeinschätzungen betrug der Mittelwert 25,5 ± 10,6 Studierende (Minimum: 6; Maximum: 69). Diese Werte entsprechen bei einer Teilnahmerate von n = 128 einer minimal notwendigen prozentualen Rücklaufquote von 19,0 % für Selbsteinschätzungen der Studierenden und 19,9 % für Fremdeinschätzungen, um reliable Ergebnisse zu erhalten. Nachdem das Vorliegen einer Normalverteilung bestätigt wurde, ergab ein t-Test zum Vergleich der benötigten Rücklaufquoten beider Methoden keinen signifikanten Unterschied (p = 0,189).

Abbildung 11 zeigt graphisch den Zusammenhang der NNE der Fremdeinschätzungs-Methode und dem entsprechenden oAPG für jedes einzelne Lernziel. Es ist zu erkennen, dass für Lernziele, die einen hohen objektiven Lernzuwachs aufwiesen, eine tendenziell geringere NNE notwendig ist als für die Lernziele, die einen niedrigen objektiven Lernzuwachs aufwiesen. Abbildung 12 zeigt ebendiesen Zusammenhang zwischen der NNE und

dem entsprechenden oAPG für jedes einzelne Lernziel für die Selbsteinschätzungs-Methode. Auch hier lässt sich erkennen, dass für Lernziele mit einem hohen oAPG tendenziell niedrigere NNE notwendig sind als für Lernziele mit einem kleinem oAPG.

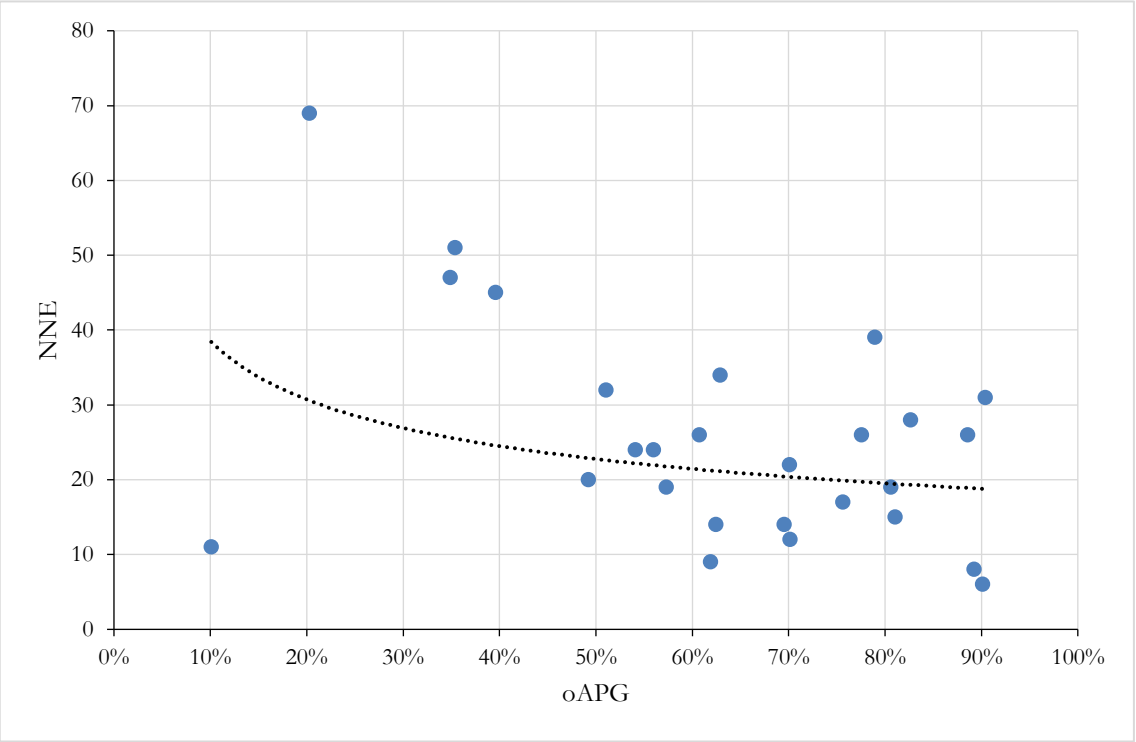

Abbildung 11: Zusammenhang zwischen der NNE der Fremdeinschätzungs-Methode und dem dazugehörigen oAPG für jedes einzelne Lernziel. Jeder Datenpunkt entspricht einem Lernziel.

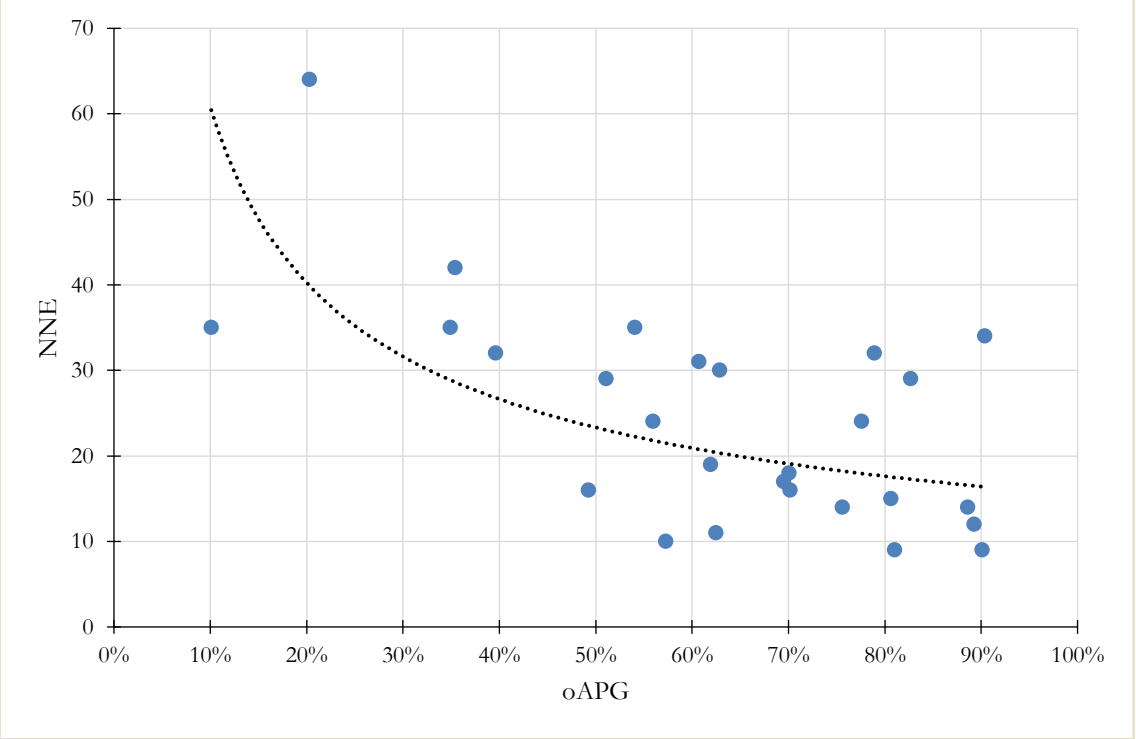

Abbildung 12: Zusammenhang zwischen der NNE der Selbsteinschätzungs-Methode und dem dazugehörigen oAPG für jedes einzelne Lernziel. Jeder Datenpunkt entspricht einem Lernziel.

Basierend auf den dargestellten Daten kann festgestellt werden, dass für die Selbsteinschätzungs-Methode und die Fremdeinschätzungs-Methode vergleichbare Rücklaufquoten erreicht werden müssen, um reliable Daten zum Lernzuwachs zu erhalten. Zusätzlich kann gefolgert werden, dass die NNE mit steigendem objektiven Lernzuwachs für beide Methoden sinkt.

3.5 Vergleich von Selbst-/Fremdeinschätzungs-Methode und objektiver Prüfung (Forschungsfrage drei)

3.5.1 Prüfungsergebnisse

Zunächst sollen die Ergebnisse der objektiven Prüfungen dargestellt werden. Zum Zeitpunkt T1 betrug der Mittelwert der erreichten Prozentzahlen der Studierenden in der objektiven Prüfung 54,7 ± 4,7 % (Minimum: 40,0 %, Maximum: 64,4 %). Zum Zeitpunkt T2 erreichten die Studierenden im Mittel 83,3 ± 9,0 % (Minimum: 54,1 %, Maximum: 94,8 %). Hierbei wurde von einer maximal möglichen Punktzahl von 135 ausgegangen, da in die Analyse nach Durchführung der Sensitivitätsanalyse nur noch 27 Items mit je fünf möglichen Punkten einflossen. Zur Analyse der internen Konsistenz wurde Cronbach's α bestimmt. Dieses betrug auf Unteritem-Ebene für die objektive Prüfung zum Zeitpunkt T1 $\alpha = 0{,}33$ und zum Zeitpunkt T2 $\alpha = 0{,}91$. Die mittlere Itemschwierigkeit betrug zum Zeitpunkt T1 0,55 ± 0,22 und reichte von 0,07 bis 0,98. Zum Zeitpunkt T2 reichte die Itemschwierigkeit von 0,05 bis

1,0 und betrug im Mittel 0,83 ± 0,18. Die mittlere Itemtrennschärfe betrug in der objektiven Prüfung zum Zeitpunkt T1 0,04 ± 0,09 sowie zum Zeitpunkt T2 0,26 ± 0,18. In der Prüfung zum Zeitpunkt T1 traten bei 45 von 135 Items negative Trennschärfen auf, in der Prüfung zu T2 lediglich bei acht Items.

3.5.2 Objektiv gemessener Lernzuwachs

Abbildung 13 zeigt die mittleren erreichten Punktzahlen der objektiven Prüfungen nach der Invertierung und Anpassung für jedes Lernziel. Die Spannweite der Mittelwerte betrug hierbei vor dem Modul 2,55 – 4,42. Das größte Vorwissen der Studierenden bestand für das Lernziel „Hypertonie“, das geringste Vorwissen für das Lernziel „Pneumonie“. Nach dem Modul betrug die Spannweite der Mittelwerte 1,15 – 2,7 mit dem größten Wissen in Bezug auf das Lernziel „Hypertonie“ und dem geringsten für das Lernziel „Hyperlipidämie“. Anschließend konnten die errechneten Mittelwerte zur Bestimmung des oAPG verwendet werden.

Beispielhaft wird hier die Berechnung des Lernzuwachses für das Lernziel „Akute Bronchitis“ in der objektiven Prüfung dargestellt:

$$\text{oAPG } [\%] = \frac{3{,}59 - 2{,}02}{3{,}59 - 1} \times 100 = 60{,}6$$

Für das Lernziel „Akute Bronchitis“ betrug der Lernzuwachs in der objektiven Prüfung demnach 60,6 %. Tabelle 3 zeigt den aggregierten Lernzuwachs in der objektiven Prüfung für jedes einzelne Lernziel. Der oAPG für spezifische Lernziele reichte von minimal 10,0 % für das Lernziel „Kardiomyopathie“ bis zu maximal 90,4 % für das Lernziel „Hypertonie“. Somit wies das Item zum Lernziel „Hypertonie“, das in der objektiven Prüfung als einziges Item durch eine Freitextfrage geprüft wurde, einerseits das größte Vorwissen im Sinne des niedrigsten Mittelwertes vor und nach dem Modul und gleichzeitig den größten Lernerfolg, im Sinne des höchsten oAPG, auf. Im Mittel war für alle 27 Lernziele ein objektiver Lernzuwachs von 63,0 ± 21,3 % zu verzeichnen.

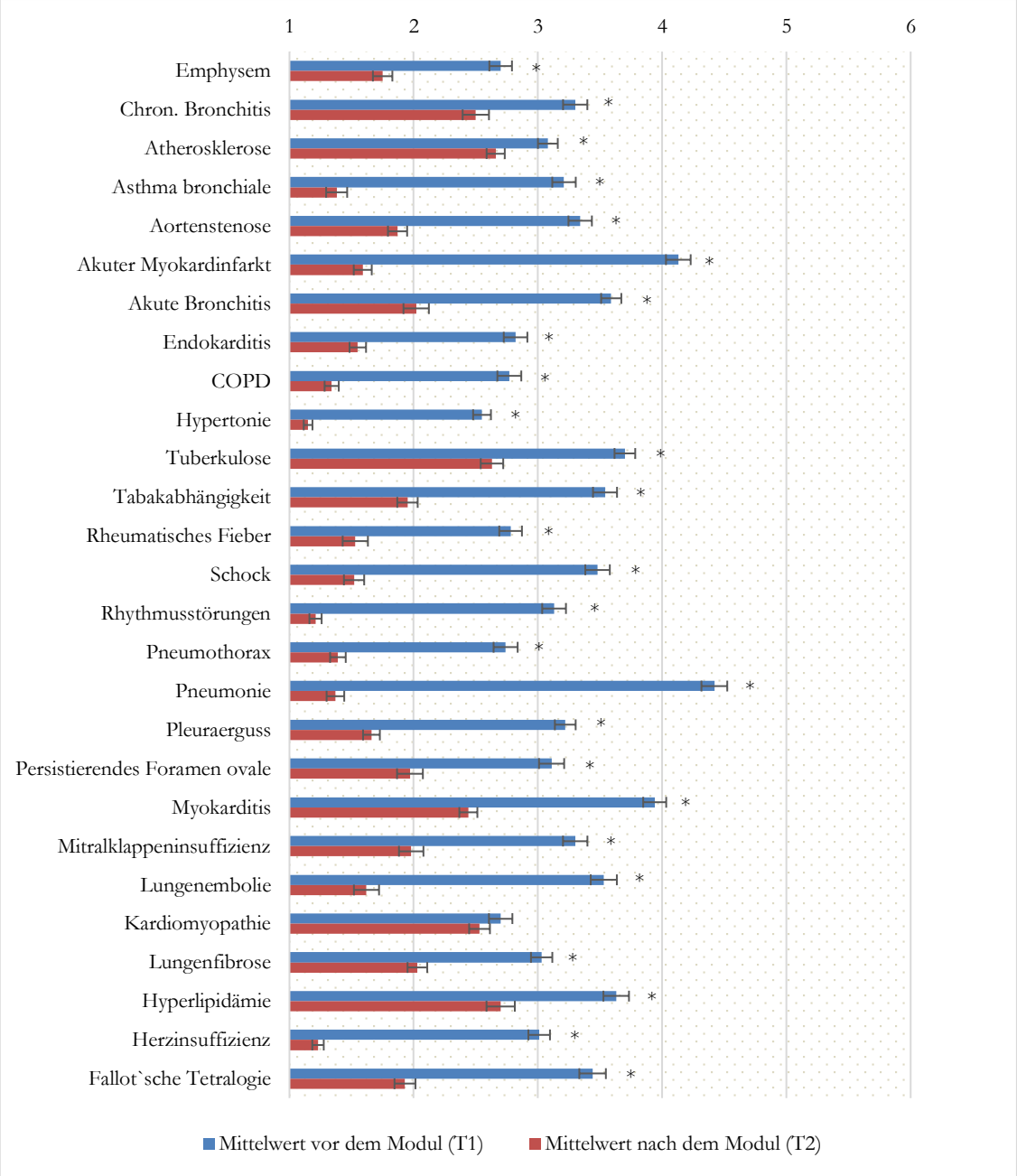

Abbildung 13: Erreichte Punktzahlen in der objektiven Prüfung für jedes Lernziel vor (T1) und nach (T2) dem Modul. Dargestellt sind die MW sowie die Standardfehler der erreichten Punktzahlen für jedes Lernziel zum Zeitpunkt T1 (in Form von blauen Balken) und T2 (in Form von roten Balken) in der objektiven Prüfung. Die Abbildung zeigt die bereits umkodierten Punktzahlen auf die sechsstufige Likert-Skala. Es konnten Werte zwischen Eins (alle Unteritems korrekt bewertet) und Sechs (kein Unteritem korrekt bewertet) erreicht werden. Zudem wurden alle Lernziele, bei denen der Unterschied der MW zwischen T1 und T2 nach Bonferroni-Korrektur im Wilcoxon-Test signifikant ausfiel mit einem „*“ markiert.

Tabelle 3: Lernzuwachs in der objektiven Prüfung in Prozent

Lernziel	Objektive Prüfung
Emphysem	55,9 %
Chronische Bronchitis	34,8 %
Atherosklerose	20,2 %
Asthma bronchiale	82,8 %
Aortenstenose	62,8 %
Akuter Myokardinfarkt	81,2 %
Akute Bronchitis	60,6 %
Endokarditis	69,8 %
COPD	80,8 %
Hypertonie	90,4 %
Tuberkulose	39,6 %
Tabakabhängigkeit	62,6 %
Rheumatisches Fieber	70,2 %
Schock	79,0 %
Rhythmusstörungen	90,1 %
Pneumothorax	77,6 %
Pneumonie	89,2 %
Pleuraerguss	70,3 %
Persistierendes Foramen ovale	54,0 %
Myokarditis	51,0 %
Mitralklappeninsuffizienz	57,4 %
Lungenembolie	75,5 %
Kardiomyopathie	10,0 %
Lungenfibrose	49,3 %
Hyperlipidämie	35,4 %
Herzinsuffizienz	88,6 %
Fallot'sche Tetralogie	61,9 %

Dargestellt ist der prozentuale Lernzuwachs in der objektiven Prüfung für jedes einzelne Lernziel.

Die Annahme einer Normalverteilung konnte für die 27 Lernziele in der objektiven Prüfung nicht bestätigt werden. Daraufhin wurde der Wilcoxon-Test für zwei verbundene Stichproben genutzt und eine Bonferroni-Korrektur durchgeführt, sodass sich ein korrigierter p-Wert von 0,00185 ergab. Es zeigte sich für 26 von 27 Lernzielen in der objektiven Prüfung ein signifikanter Unterschied der Mittelwerte zwischen T1 und T2. Für das Lernziel „Kardiomyopathie" ergab sich kein signifikanter Unterschied. Ob es sich für jedes einzelne Lernziel um eine signifikante Veränderung handelt ist in Abbildung 13 gekennzeichnet.

3.5.3 Korrelationen

Im folgenden Abschnitt soll Forschungsfrage drei (Wie hängt der aus Selbsteinschätzungen und Fremdeinschätzungen errechnete Lernerfolg mit dem in wiederholten Prüfungen erhobenen objektiven Lernerfolg zusammen?) beantwortet werden. Da sowohl für die objektive Prüfung als auch für die Selbsteinschätzungen und Fremdeinschätzungen Prozentwerte für

den mittleren aggregierten Lernzuwachs berechnet wurden, konnte untersucht werden, inwiefern die Lernzuwachs-Daten miteinander korrelieren. Abbildung 14 zeigt den Lernzuwachs laut Selbsteinschätzungen, Fremdeinschätzungen und objektiver Prüfung im direkten Vergleich.

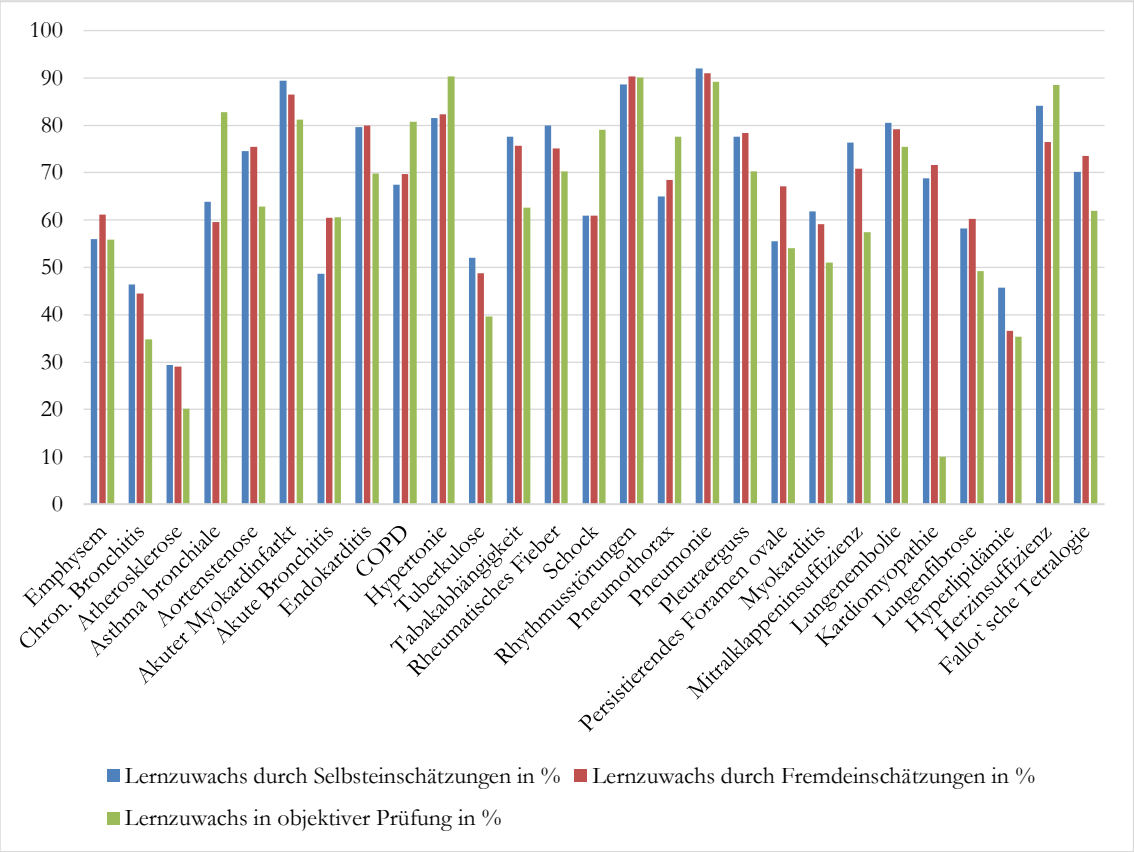

Abbildung 14: Vergleich des Lernzuwachses aus Selbsteinschätzungen, Fremdeinschätzungen und objektiver Prüfung. Die Abbildung zeigt den Lernzuwachs in Prozent für jedes einzelne Lernziel. Die blauen Balken entsprechen den sAPG-Werten, die roten den pAPG-Werten und die grünen den oAPG-Werten.

Die Korrelation zwischen dem sAPG und dem oAPG ($r = 0{,}709$; $p < 0{,}001$) ist in Abbildung 15 visualisiert, die Korrelation zwischen dem pAPG und dem oAPG ($r = 0{,}704$; $p < 0{,}001$) in Abbildung 16.

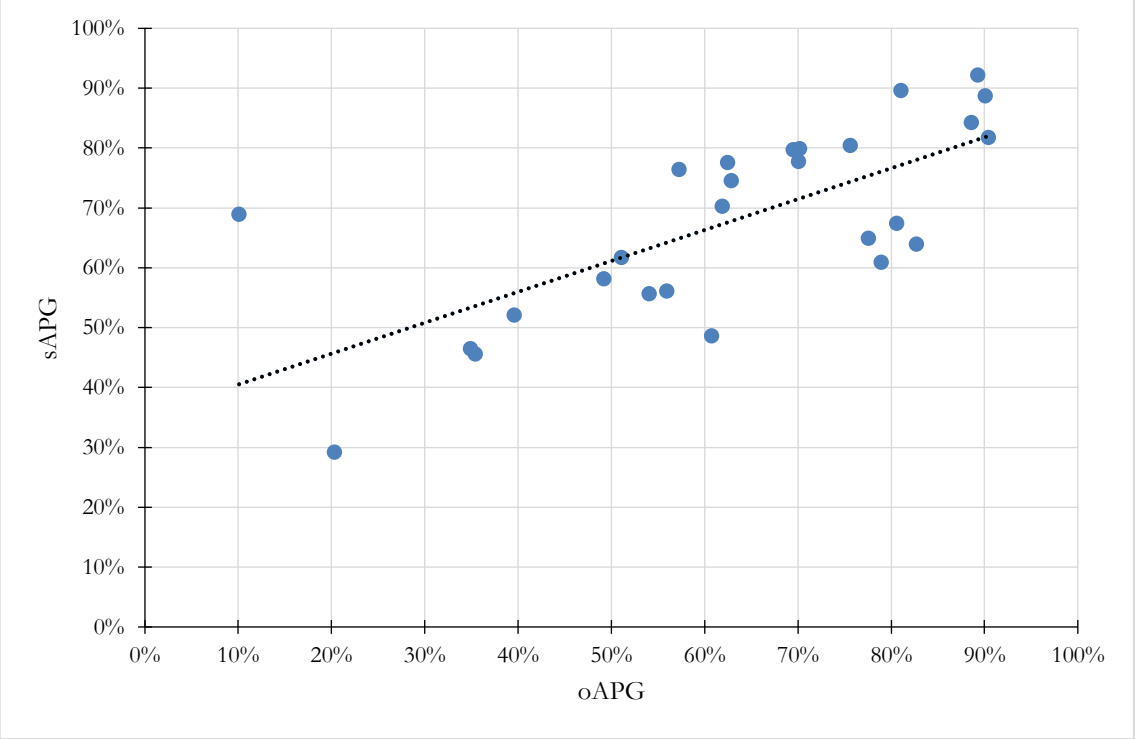

Abbildung 15: Korrelation zwischen dem Lernzuwachs aus Selbsteinschätzungen und objektiver Prüfung. Jeder einzelne Datenpunkt entspricht dabei einem Lernziel.

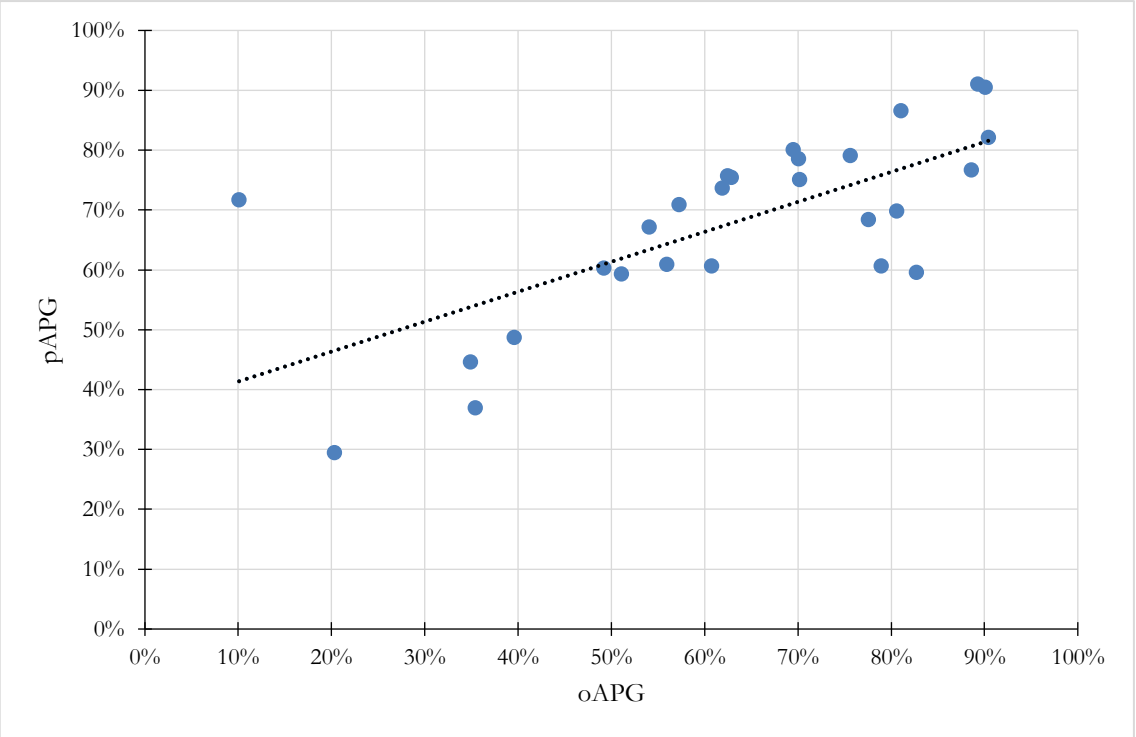

Abbildung 16: Korrelation zwischen dem Lernzuwachs aus Fremdeinschätzungen und objektiver Prüfung. Jeder einzelne Datenpunkt entspricht dabei einem Lernziel.

Es kann damit festgestellt werden, dass ein hoher signifikanter Zusammenhang zwischen Selbsteinschätzungs- bzw. Fremdeinschätzungs-Methode und objektiver Prüfung besteht. Allerdings ist die Korrelation in beiden Fällen schwächer als die Korrelation zwischen dem

sAPG und dem pAPG. Das liegt vor allem am Item zum Lernziel „Kardiomyopathie“. Wie in Kapitel 3.2 bereits erläutert ist dieses eines der fünf Items, das eine starke Abweichung zwischen den Aussagen der Selbst- und Fremdeinschätzungen und der objektiven Prüfung zeigt. In Abbildung 14 lässt sich erkennen, dass der Lernzuwachs der objektiven Prüfung für dieses Lernziel geringer ist als der Lernzuwachs für das gleiche Lernziel aus Selbst- und Fremdeinschätzungen. Diese Besonderheit ist auch in den Abbildung 15 und Abbildung 16 identifizierbar.

Es wurde zudem ein Bland-Altman-Plot erstellt, der den Zusammenhang zwischen dem pAPG und dem oAPG darstellt (siehe Abbildung 17). Der Mittelwert der Differenz aus dem prozentualen Lernzuwachs von beiden Methoden betrug 4,9 %, wobei der pAPG im Mittel höher ausfiel als der oAPG. In Abbildung 17 lässt sich zudem erkennen, dass bis auf ein Lernziel alle anderen innerhalb des 95 % *limits of agreement* (entspricht: Mittelwert der Differenz aus beiden Methoden ± 1,96 × Standardabweichung) liegen und somit ein hoher Zusammenhang besteht. Das Item außerhalb dieses Bereiches gehört zum Lernziel „Kardiomyopathie“. Dieses wies, wie bereits zuvor beschrieben, eine große Differenz zwischen dem Lernzuwachs aus der objektiven Prüfung und dem Lernzuwachs aus den Selbst- und Fremdeinschätzungen auf.

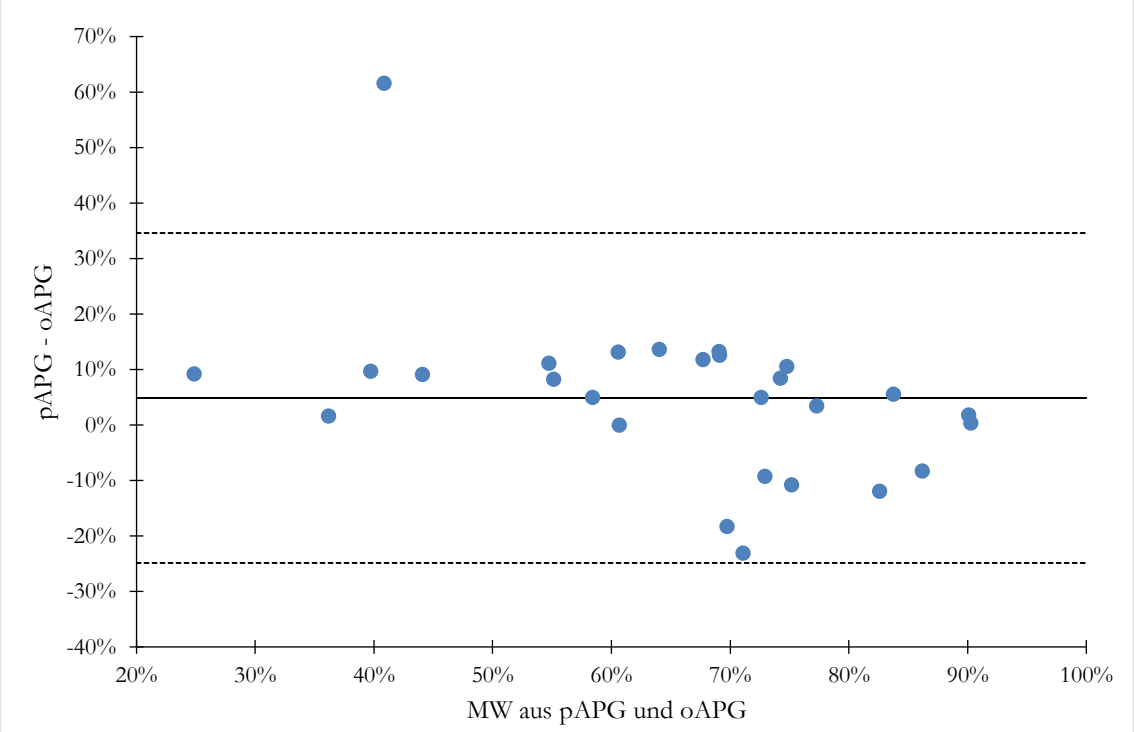

Abbildung 17: Bland-Altman-Plot des Lernzuwachses aus Fremdeinschätzungen und objektiver Prüfung. Jeder Datenpunkt entspricht einem Lernziel. Zudem sind drei waagerechte Hilfslinien eingezeichnet. Die durchgezogene Linie entspricht dem MW der Lernzuwachs-Differenzen aller 27 Lernziele. Die gestrichelten Linien entsprechen diesem MW ± 1,96 × Standardabweichung des MW der Differenz.

Abbildung 18 zeigt einen Bland-Altman-Plot, der den Zusammenhang zwischen sAPG und oAPG abbildet. Die mittlere Differenz zwischen sAPG und oAPG betrug 4,9 %. Auch der sAPG fiel im Mittel höher aus als der oAPG. Wie schon im vorherigen Bland-Altman-Plot, der den Zusammenhang zwischen pAPG und oAPG zeigt, liegen auch in diesem Fall abgesehen von einem Lernziel alle anderen innerhalb des 95 % *limits of agreement*. Das Item außerhalb dieses Bereichs bildet auch in diesem Fall das Lernziel „Kardiomyopathie" ab. Es besteht somit auch zwischen den Werten von sAPG und oAPG ein starker Zusammenhang.

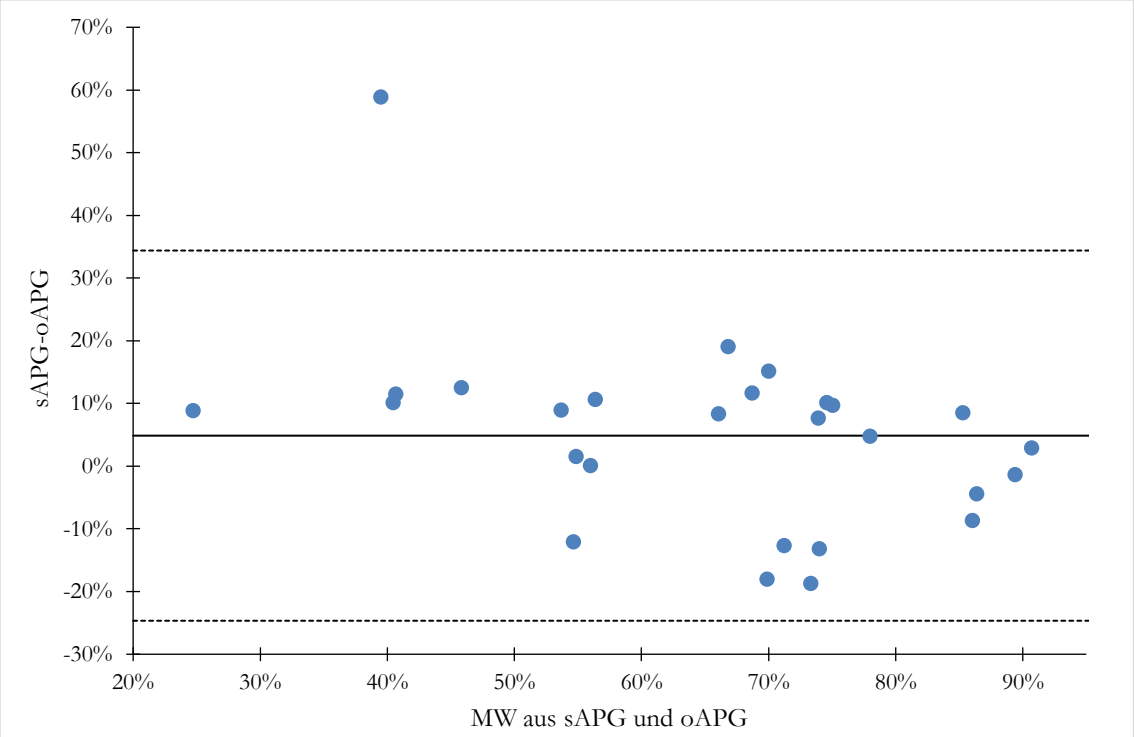

Abbildung 18: Bland-Altman-Plot des Lernzuwachses aus Selbsteinschätzungen und objektiver Prüfung. Jeder Datenpunkt entspricht einem Lernziel. Zudem sind drei waagerechte Hilfslinien eingezeichnet. Die durchgezogene Linie entspricht dem MW der Lernzuwachs-Differenzen aller 27 Lernziele. Die gestrichelten Linien entsprechen diesem MW ± 1,96 × Standardabweichung des MW der Differenz.

Auf Grundlage der dargestellten Daten konnte ein starker Zusammenhang zwischen dem auf Selbsteinschätzungen bzw. Fremdeinschätzungen basierenden Lernzuwachs und dem Lernzuwachs in einer objektiven Prüfung gezeigt werden. Die Korrelation zwischen Selbsteinschätzungs- bzw. Fremdeinschätzungs-Methode und objektiver Prüfung war in beiden Fällen signifikant ($p < 0{,}001$). Es kann somit gefolgert werden, dass beide Methoden geeignet sind, um den Lernzuwachs innerhalb eines Moduls adäquat widerzuspiegeln.

3.6 *Receiver-operating-characteristics*-Analyse

Zur Festlegung eines *Cut-off*-Wertes, der die Bestimmung von optimalem und suboptimalem objektivem Lernzuwachs auf dem Boden von Selbst- bzw. Fremdeinschätzungen erlaubt, wurde eine ROC-Analyse durchgeführt (siehe Abbildung 19). Es wurden verschiedene *Cut-off*-Werte für die Selbst- und Fremdeinschätzungs-Methode getestet und hierbei der Wert

bestimmt, der die beste Sensitivität bei höchster Spezifität ergab. Die Analyse ergab sowohl für die Selbsteinschätzungs-Methode als auch für die Fremdeinschätzungs-Methode einen optimalen *Cut-off*-Wert von 61 %. Für die Selbsteinschätzungs-Methode ergab sich bei diesem *Cut-off*-Wert eine Sensitivität von 86 % und eine Spezifität von 83 %. Der PPW lag bei 95 %, der NPW bei 62 %. Die Fremdeinschätzungs-Methode zeigte bei gleichem *Cut-off*-Wert eine Sensitivität von 90 %, eine Spezifität von 83 % sowie einen PPW von 95 % und einen NPW von 71 %.

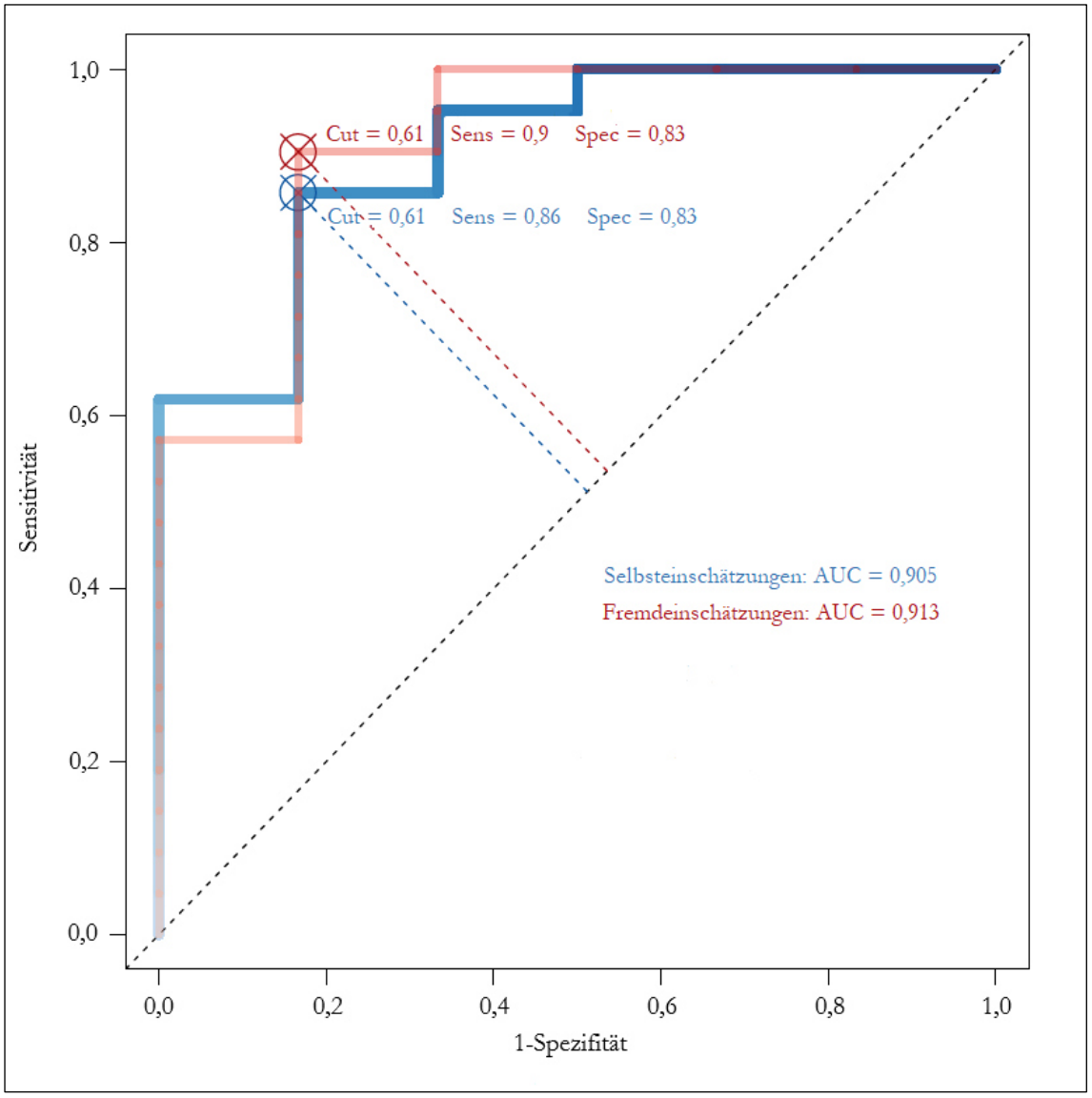

Abbildung 19: ROC-Analyse. Die Abbildung zeigt die ROC-Kurven sowohl für die Selbsteinschätzungs-Methode (blaue Kurve) als auch für die Fremdeinschätzungs-Methode (rote Kurve). Es sind die *Cut-off*-Werte (*Cut*), die Sensitivität (Sens), die Spezifität (Spec) sowie die *area under the curve* (AUC) angegeben.

4 Diskussion

Im Rahmen der vorliegenden Arbeit wurde eine prospektive Kohortenstudie durchgeführt, die der Weiterentwicklung des an der UMG implementierten Evaluationssystems dienen soll. Es wurde untersucht, ob die Vorhersage-basierte Evaluationsmethode auf das Instrument der Lernerfolgsevaluation mittels spezifischer Lernziele übertragbar ist. Außerdem wurde analysiert, wie die Ergebnisse dieser Methode mit den Ergebnissen der ursprünglich genutzten Methode der vergleichenden studentischen Selbsteinschätzungen zur Berechnung des lernzielbezogenen Lernerfolgs zusammenhängen. Zusätzlich wurde geprüft, wie die Lernzuwachsraten beider Methoden mit den objektiv erhobenen Lernzuwachsraten zusammenhängen. Des Weiteren wurden die minimal notwendigen Rücklaufquoten für beide Methoden bestimmt, um reliable Ergebnisse zu erhalten.

Bezogen auf die in Kapitel 1.4 vorgestellten Forschungsfragen ergeben sich folgende Beobachtungen:

Die Selbsteinschätzungs- und die Fremdeinschätzungs-Methode lieferten vergleichbare Lernerfolgsdaten, die einen starken, signifikanten Zusammenhang aufwiesen (Forschungsfrage eins).

Beide Methoden lieferten vergleichbare minimal notwendige Rücklaufquoten, um reliable Lernerfolgsdaten zu generieren (knapp 20 %). Zudem zeigte sich für beide Methoden, dass Lernziele mit einem hohen objektiven Lernzuwachs niedrigere Rücklaufquoten erforderten als Lernziele mit einem geringen objektiven Lernzuwachs (Forschungsfrage zwei).

Die mit beiden Einschätzungsmethoden gewonnenen Lernerfolgsdaten wiesen signifikante Korrelationen mit dem Lernzuwachs in der objektiven Prüfung auf (Forschungsfrage drei).

4.1 Zusammenhang zwischen Selbsteinschätzungs- und Fremdeinschätzungs-Methode (Forschungsfrage eins)

Die Nutzung von studentischen Selbsteinschätzungen zur Messung des Einflusses der Lehre ist weit verbreitet (Drennan und Hyde 2008). Zudem ist das Konzept der vergleichenden studentischen Selbsteinschätzungen ein valides Instrument zur Bestimmung des studentischen Lernerfolgs (Raupach et al. 2011; Raupach et al. 2012; Schiekirka et al. 2013). Die in dieser Studie untersuchte Fremdeinschätzungs-Methode stellt hingegen eine vergleichsweise neue Herangehensweise dar. Die vorliegende Studie basierte auf dem Prinzip der Vorhersage-basierten Evaluationsmethode und verband die Fremdeinschätzungs-Methode mit dem Prinzip der Lernziel-basierten Evaluation. Während erste Studien bereits den Einsatz der Vorhersage-basierten Evaluationsmethode zur Lehrevaluation im Medizinstudium testeten (Cohen-Schotanus et al. 2010; Schönrock-Adema et al. 2013), war dies die erste Studie, welche die Anwendung auf die Lernerfolgsevaluation untersuchte.

Im Rahmen der vorliegenden Forschungsarbeit wurde untersucht, ob die Vorhersage-basierte Methode, die hier als Fremdeinschätzungs-Methode bezeichnet wird, vergleichbare Lernerfolgsdaten liefert wie die Selbsteinschätzungs-Methode und ob sich durch die Erhebung von Fremdeinschätzungen der erforderliche Mindestrücklauf reduzieren lässt. Die Ergebnisse zeigten, dass die Fremdeinschätzungs-Methode gleichwertige Lernerfolgsdaten liefert wie die Selbsteinschätzungs-Methode.

Lediglich ein Lernziel wies in der vorliegenden Studie einen Lernzuwachs von circa 30 % für beide Evaluationsmethoden auf. Die anderen 26 Lernziele ergaben für beide Methoden einen höheren Lernzuwachs. Da unterhalb dieser Grenze keine weiteren Daten vorlagen, kann keine zuverlässige Aussage über die Vergleichbarkeit beider Methoden bei niedrigen Lernzuwachsraten (< 30 %) getroffen werden. Oberhalb der Grenze von 30 % besteht jedoch eine sehr gute Vergleichbarkeit beider Methoden. Es kann gefolgert werden, dass die Fremdeinschätzungs-Methode für Lernziele mit einem Lernzuwachs von mindestens 30 % gleichwertig zur Selbsteinschätzungs-Methode ist, was die Nutzung als Instrument zur Bestimmung des vorhergesagten bzw. selbsteingeschätzten studentischen Lernzuwachses, bezogen auf spezifische Lernziele, betrifft.

Im Gegensatz zu bisherigen Studien, welche die Vorhersage-basierte Evaluationsmethode untersuchten, konnten in der vorliegenden Studie einige Einschränkungen vermieden werden. In Studien von Cohen-Schotanus et al. (2010) und Schönrock-Adema et al. (2013) wurden zur Erhebung der persönlichen Bewertungen der Studierenden vierstufige Likert-Skalen genutzt. Dabei sollten die Studierenden sich für die ihrer Meinung nach zutreffendste der vier Optionen entscheiden und diese ankreuzen. Diejenigen Studierenden, welche in den beschriebenen Studien um die Erhebung der Vorhersagen gebeten wurden, mussten im Gegensatz dazu nicht eine Option auf einer vierstufigen Likert-Skala ankreuzen. Stattdessen wurden ihnen zur Erhebung der Vorhersagen pro Frage vier Antwortkategorien präsentiert, die den vier Antwortmöglichkeiten der Likert-Skalen zur Erhebung der persönlichen Bewertungen entsprachen. Für jede der Antwortkategorien existierte dabei ein Freitextfeld, in das die Studierenden ihre Vorhersagen darüber eintragen sollten, wieviel Prozent ihrer Kommilitonen sich im Rahmen der Erhebung der persönlichen Bewertungen für diese Antwortkategorie der vierstufigen Likert-Skala entscheiden würden. Somit mussten zur Erhebung der Vorhersagen pro Frage vier Prozentzahlen auf vier Antwortkategorien aufgeteilt werden. Allerdings erwies sich diese Methode als störanfällig. Einige der Studierenden hatten Probleme, die Prozentzahlen so auf die vier Kategorien aufzuteilen, dass die Summe aller Kategorien genau 100 % ergab. Zudem tendierten die Studierenden dazu, „runde" Zahlen in Schritten von 5 – 10 % über die vier Kategorien zu verteilen und alle vier Kategorien der Skala zu verwenden (Cohen-Schotanus et al. 2010). Durch die beschriebene Weise der Erhebung von persönlichen Bewertungen und Vorhersagen waren die beiden Erhebungsinstrumente somit nicht vollständig kongruent. Im Gegensatz dazu wurde in der vorliegenden Studie eine identische, sechsstufige Likert-Skala zur Erhebung der Selbst- und Fremdeinschätzungen genutzt. Hierdurch kam es nicht zur fehlerhaften Berechnung von

Summen oder Tendenzen in der Verteilung der Prozentzahlen. Die Selbst- und Fremdeinschätzungsaussagen konnten somit durch Nutzung der gleichen Likert-Skala exakt miteinander verglichen werden, was eine Stärke im Vergleich zu bisherigen Studien darstellt.

Eine weitere Problematik, die in der Literatur beschrieben wurde, bestand darin, dass keine objektiven Messwerte zur Überprüfung der Exaktheit der Ergebnisse erhoben wurden (Cohen-Schotanus et al. 2010). In der vorliegenden Studie wurde hingegen durch die Erhebung von objektiven Werten zum Lernzuwachs eine weitere Dimension beachtet, die den direkten Vergleich der Selbst- und Fremdeinschätzungen mit den objektiven Daten erlaubt. Die Ergebnisse werden in Kapitel 4.3 tiefergehend diskutiert.

4.2 Rücklaufquoten (Forschungsfrage zwei)

Durch eine geringe Teilnahmerate an Evaluationsprozessen wird die Validität der genutzten Instrumente in Frage gestellt (Cummings et al. 2001; Kellerman 2001; VanGeest et al. 2007). Je höher hingegen die Teilnahmerate ist, desto mehr kann den Evaluationsergebnissen vertraut werden (O'Rourke 1999). Es sollte demnach festgelegt werden können, welche Teilnahmerate notwendig bzw. ausreichend ist, um verlässliche Schlüsse aus den Evaluationen ziehen zu können. In der vorliegenden Studie wurden die minimal notwendigen Rücklaufquoten der Selbst- und Fremdeinschätzungs-Methode zu jeweils knapp 20 % bestimmt.

Im Kontext globaler Kursbewertungen unterscheidet sich die notwendige Rücklaufquote zwischen der Vorhersage-basierten Evaluationsmethode und der Methode basierend auf persönlichen Bewertungen der Studierenden signifikant, wobei die Vorhersage-basierte Methode einen signifikant niedrigeren Rücklauf an Evaluationen erfordert (Cohen-Schotanus et al. 2010). In der vorliegenden Studie konnte diese Beobachtung bezüglich des Lernerfolgs nicht gemacht werden. Laut Cohen-Schotanus et al. (2010) liegen generelle Empfehlungen zum Rücklauf von Evaluationen bei 65 – 70 % und auch die notwendige Rücklaufquote der Methode basierend auf persönlichen Bewertungen ordnen sie in dieser Größenordnung ein. Für die Vorhersage-basierte Methode lag die minimal notwendige Zahl an Studierenden der Literatur zufolge zwischen 20 und 40, je nachdem ob die Studierenden vorab Informationen zu vorangegangenen Evaluationsergebnissen erhielten oder nicht (Cohen-Schotanus et al. 2010). Eine andere Vorstudie der gleichen Arbeitsgruppe bestätigte die signifikant niedrigeren, notwendigen Rücklaufquoten bei der Nutzung von Vorhersagen, im Gegensatz zu persönlichen Bewertungen der Studierenden. Hier zeigte sich, dass für die eigenen Bewertungen zwischen 67 – 79 Studierende (bei n = 106 – 157 Teilnehmenden) notwendig waren. Dies würde einer Rücklaufquote, wie sie in dieser Studie berechnet wurde, von 45,4 – 66,3 % entsprechen. Für die Vorhersage-basierte Methode waren hingegen nur 26 – 28 Studierende notwendig (bei einer Teilnahmerate von n = 92 – 122), was einer Rücklaufquote von 22,9 – 28,5 % wie in dieser Studie berechnet, entsprechen würde (Schönrock-Adema et al. 2013). Zusammenfassend folgerten die Autoren, dass eine Kohorte von 30 – 40 Studierenden ausreichend ist, um die Meinung dieser Studienkohorte adäquat zu erfassen

(Schönrock-Adema et al. 2013). Die vorliegende Studie ergab für die Erhebung reliabler Lernerfolgsdaten insgesamt niedrigere notwendige Rücklaufquoten für beide Methoden ohne wesentlichen Unterschied zwischen Selbst- und Fremdeinschätzungen. Während in der bisherigen Forschung die Vorhersage-basierte Methode zur Lehrevaluation deutlich niedrigere Rücklaufquoten erforderte als die Nutzung von eigenen Bewertungen der Studierenden über den Kurs, konnte in der vorgestellten Studie keine Reduzierung des notwendigen Rücklaufs durch diese neue Methode erzielt werden. Eine mögliche Erklärung hierfür ist, dass in der vorliegenden Studie ein anderer Zielparameter gewählt wurde. Studentische Bewertungen der Lehre werden durch verschiedene Arten von *bias* beeinflusst (siehe Kapitel 1.3.2.1). Die beschriebenen *bias* wirken sich stärker auf Evaluationsergebnisse aus, wenn Studierende nach ihrer eigenen Zufriedenheit mit der Lehre anstatt nach ihrem Kompetenzgewinn gefragt werden (Braun und Leidner 2009), der in diesem vorliegenden Fall dem Lernzuwachs entspricht. In der Literatur zeigte sich eine deutliche Verringerung der notwendigen Rücklaufquote, wenn die Studierenden nicht ihre eigene Zufriedenheit oder Meinung, sondern die ihrer Kommilitonen angeben sollten, da hierdurch die persönlichen Einflüsse minimiert wurden (Cohen-Schotanus et al. 2010; Schönrock-Adema et al. 2013). Im Gegensatz dazu wurde in der vorliegenden Studie der Lernzuwachs der Studierenden erhoben. Es lässt sich vermuten, dass die Einschätzung dieses Lernzuwachses nicht in gleichem Maße durch persönliche Einflüsse und *bias* verändert wird, wie die Zufriedenheit oder Bewertung eines Kurses und aus diesem Grund der Unterschied, der sich in der Erhebung der Selbst- und Fremdeinschätzungen zeigte, nur sehr gering war. Zudem lässt sich anhand der Ergebnisse aus der vorliegenden Studie vermuten, dass Studierende die Zufriedenheit ihrer Kommilitonen mit einem Kurs und die Globalbewertungen ihrer Kommilitonen über ebendiesen Kurs deutlich besser einschätzen können, als den Leistungsstand ihrer Kommilitonen zu spezifischen Lernzielen.

In der vorgestellten Studie konnte für beide Evaluationsmethoden ein Zusammenhang zwischen dem notwendigen Evaluationsrücklauf und dem objektiven Lernzuwachs festgestellt werden. Je höher das oAPG für ein spezifisches Lernziel ausfiel, desto niedriger war die NNE. Diese Beobachtung zeigte sich sowohl für die Selbst- als auch für die Fremdeinschätzungs-Methode.

Dieser Zusammenhang zwischen dem oAPG und der dazugehörigen NNE eröffnet eine neue Dimension in der Nutzung der Evaluationsmethoden mittels vergleichender Selbst- und Fremdeinschätzungen. Durch die Erhebung von objektiv gemessenen Daten zum Lernzuwachs in der vorliegenden Studie ist ein Vergleich der Selbst- sowie der Fremdeinschätzungs-Methode mit objektiven Kriterien möglich. Hierdurch kann auch eine genauere Differenzierung bezüglich der Zusammenhänge in Abhängigkeit von der Höhe des tatsächlichen, objektiven Lernzuwachses erfolgen. Die Schlussfolgerungen hierzu sind in Kapitel 4.5.1 beschrieben.

4.3 Vergleich der Lernzuwachsraten beider Evaluationsmethoden mit dem Lernzuwachs der objektiven Prüfung (Forschungsfrage drei)

Die dritte Forschungsfrage beschäftigte sich mit der Thematik des Zusammenhangs zwischen dem sAPG und oAPG sowie dem pAPG und oAPG. Die in beiden Fällen hohe Korrelation erlaubt die Schlussfolgerung, dass beide Methoden geeignet sind, um den objektiven Lernzuwachs innerhalb eines Moduls adäquat widerzuspiegeln (Forschungsfrage drei).

Der mittlere objektive Lernzuwachs betrug 63,0 ± 21,3 % und lag somit etwas niedriger als sAPG und pAPG (jeweils knapp 68 %). Dabei muss beachtet werden, dass der Großteil der objektiven Prüfung aus Fragen, die sich aus mehreren *True-False-Items* zusammensetzten, konzipiert wurde. Hierbei bestand jede einzelne Aufgabe aus einem Aufgabenstamm, dem fünf Aussagen folgten, die von den Studierenden als „richtig" oder „falsch" markiert werden sollten (Albanese et al. 1979). Für diese Art der Aufgabenstellungen, die auch als KPrim-Fragen bezeichnet werden, existiert eine Reihe von verschiedenen Möglichkeiten zur Bewertung (Kanzow et al. 2018). Kanzow et al. (2018) beschrieben ausführlich die Auswirkungen auf die erzielten Punktzahlen durch die Nutzung unterschiedlicher Bewertungsmethoden. In der vorliegenden Studie wurden die Fragen anhand der *Multiple-True-False*-Methode bewertet. Hierbei wurde jede richtige Antwort mit einem Punkt bewertet (Albanese et al. 1979). Somit konnte für jedes Unteritem entweder ein Punkt erreicht werden, wenn die korrekte Antwort ausgewählt wurde oder kein Punkt, wenn die falsche Antwortmöglichkeit gewählt wurde. Die erreichte Punktzahl pro Aufgabe (maximal mögliche Punktzahl: fünf) ergab sich aus der Summe der Punktzahlen der Unteritems einer Frage und die Gesamtpunktzahl der Prüfung (maximal mögliche Punktzahl: 135) somit aus der Summe der Punktzahlen aller Unteritems. Für jedes einzelne Unteritem innerhalb der Frage bestand eine Ratewahrscheinlichkeit von 50 %. Durch die Nutzung dieser Bewertungsmethode wurde jedoch auch ein Teilwissen durch erzielte Punkte belohnt und in der erzielten Punktzahl spiegelten sich 100 % der Informationen aus der Aufgabestellung wider (Kanzow et al. 2018). Somit bestand in der vorliegenden Studie jedoch auch für die Gesamtprüfung eine Ratewahrscheinlichkeit von 50 %. Diese Wahrscheinlichkeit spiegelt sich in etwa in dem mittleren Prüfungsergebnis von 54,7 ± 4,7 % zu T1 wider und ist in Abbildung 13 zu erkennen. Hier stellt sich der Einfluss der Ratewahrscheinlichkeit von 50 % pro Unteritem dar, indem die Mittelwerte der Lernziele in der objektiven Prüfung zu T1 für die meisten Lernziele nahe um den Wert Drei, entsprechend 50 % der möglichen Punkte pro Lernziel, streuen. Es kann davon ausgegangen werden, dass die Studierenden einen Großteil der Fragen, die sie an diesem Zeitpunkt richtig beantwortet haben, lediglich zufällig richtig beantwortet haben. Trotz der hohen Ratewahrscheinlichkeit wurde in der vorliegenden Studie die Bewertung eines jeden der fünf Unteritems pro Frage mit null Punkten oder einem Punkt gewählt, sodass durch diese Methode keine Informationen zum Lernzuwachs verloren gingen und der Lernzuwachs genau berechnet werden konnte. Durch die Vergabe von Leistungspunkten erst oberhalb der Grenze von 50 % aller möglichen Punkte (siehe Kapitel 2.1.2) wurden die Ergebnisse für diese adjustiert.

Das Item zum Lernziel „Hypertonie“ wurde in den objektiven Prüfungen als einziges Item anhand einer Freitextfrage geprüft. Dieses wies einerseits das größte Vorwissen der Studierenden zu T1 und auch das größte Wissen zu T2, sowie zudem den größten Wert für den oAPG von 90,4 % auf. Die Itemschwierigkeiten dieses Items betrugen dabei 0,69 zu T1 und 0,97 zu T2. Somit war es insgesamt an beiden Zeitpunkten leichter zu beantworten als der Durchschnitt aller Items der objektiven Prüfungen. Es lässt sich vermuten, dass dieses Item, obwohl es im Gegensatz zu den übrigen *True-False-Item*-Fragen keine Ratewahrscheinlichkeit von 50 % aufwies, für die Studierenden nicht schwieriger zu beantworten war.

Zum Zeitpunkt T2 lag die Erfolgsquote in der objektiven Prüfung bei 83,3 ± 9,0 %. Da die Prüfung zu diesem Zeitpunkt exakt die gleichen Aufgaben enthielt wie die objektive Prüfung zum Zeitpunkt T1, kann nicht sicher ausgeschlossen werden, dass die Studierenden sich teilweise an die Aufgabenstellungen erinnerten. Allerdings wurden die Studierenden im Vorhinein nicht darüber informiert, dass sich die Prüfungen zu beiden Zeitpunkten exakt gleichen würden. Außerdem wurden nach der objektiven Prüfung zu T1 weder die Aufgabenstellungen noch die Lösungen der Aufgaben veröffentlicht. Des Weiteren lagen mehrere Wochen zwischen den zwei Prüfungszeitpunkten, sodass ein vollständiges Erinnern der Prüfungsaufgaben unwahrscheinlich bleibt. Ein Vorteil dieser Vorgehensweise besteht daher in der vollständigen Vergleichbarkeit der Prüfungsergebnisse beider objektiven Prüfungen. Zudem war die Prüfung zu T2, im Gegensatz zu derjenigen zu T1, summativer Natur. Summative Prüfungen haben dabei einen bewertenden Charakter, und werden häufig als Grundlage für die Vergabe von Noten genutzt (Roediger und Karpicke 2006). In der vorliegenden Studie konnten zu T2 bis zu 20 % der Leistungspunkte für das Modul 3.1 erworben werden, wodurch diese Prüfung summativer Natur war. Da summative Prüfungen für Studierende einen der stärksten Anreize zum Lernen für eine Prüfung darstellen (Raupach et al. 2013), kann davon ausgegangen werden, dass auch in der vorliegenden Studie der summative Charakter der Prüfung zu T2 zu einer Veränderung im Lernverhalten der Studierenden geführt hat. Die Möglichkeit zum Erwerb der Leistungspunkte hat dabei wahrscheinlich einen Anreiz für die Studierenden dargestellt, ein gutes Ergebnisses erzielen zu wollen. Raupach et al. (2013; 2016) konnten zudem zeigen, dass summative Prüfungen mit einer besseren Prüfungsleistung der Studierenden assoziiert sind als formative Prüfungen. Zusammenfassend kann davon ausgegangen werden, dass es sich bei der Differenz der Erfolgsquoten an beiden Zeitpunkten um den tatsächlichen, objektiven Lernzuwachs handelt.

Dabei zeigte die objektive Prüfung zum Zeitpunkt T1 lediglich einen Wert von 0,33 für Cronbach‘s α, während derjenige zu T2 bei $\alpha = 0{,}91$ lag. Cronbach‘s α ist ein häufig genutztes Maß zur Bestimmung der Reliabilität eines Messinstrumentes und kann Werte zwischen Null und Eins annehmen (Gliem und Gliem 2003; Tavakol und Dennick 2011). Die interne Konsistenz beschreibt dabei, inwieweit durch die Gesamtheit der Items eines Instrumentes das gleiche Konstrukt gemessen wird (Tavakol und Dennick 2011). Hierbei gelten Werte zwischen 0,7 und 0,8 als zufriedenstellend (Bland und Altman 1997). Die Prüfung zu T2 wies demnach eine hohe, diejenige zu T1 eine niedrige interne Konsistenz auf, wobei in beiden

Fällen die exakt gleiche Prüfung verwendet wurde. Mögliche Erklärungen für einen niedrigen Wert sind eine geringe Anzahl an Items, eine niedriger Zusammenhang der einzelnen Items oder heterogene Konstrukte, die diesen zugrunde liegen (Tavakol und Dennick 2011). In der vorliegenden Studie zeigten die einzelnen Mittelwerte der 135 Unteritems zu T2 ein homogeneres Antwortverhalten der Studierenden, während zum Zeitpunkt T1 die Streuung im Antwortverhalten der Studierenden größer war. Eine denkbare Erklärung hierfür ist, dass die Studierenden in der objektiven Prüfung zu T1 fast ausschließlich geraten haben (passend zur Erfolgsquote von 54,7 ± 4,7 %), wobei diesen geratenen Antworten kein einheitliches Konstrukt zugrunde liegt. Zum Zeitpunkt T2 war das Wissen der Studierenden in Bezug auf die gleichen Fragen deutlich größer (passend zur Erfolgsquote von 83,3 ± 9,0 %), wobei diesem Wissen vermutlich ein einheitlicheres Konstrukt zugrunde liegt. Dieses resultiert folglich in einer geringeren Streuung der Mittelwerte über alle Unteritems und dadurch in einer höheren internen Konsistenz der Prüfung zum Zeitpunkt T2. Aufgrund dessen und der Tatsache, dass in der vorliegenden Studie die Skala nur zwei Antwortmöglichkeiten (die Werte Null oder Eins) enthielt, ergeben sich hierdurch schnell größere Schwankungen in den Werten für Cronbach's α. Diese Ansätze stellen eine denkbare Erklärung für die Differenz der Werte für Cronbach's α zu T1 und T2 dar, eine abschließende Klärung ist hierdurch jedoch nicht sicher möglich.

4.4 Stärken und Schwächen

Im Folgenden sollen Stärken und Schwächen der vorgestellten Studie beschrieben werden.

4.4.1 Studiendesign

Die Erhebung der Daten für die vorliegende Studie erfolgte in einem einzelnen Modul an einer medizinischen Fakultät in Deutschland. Da in der Studie lediglich die Lehre im Modul 3.1 an der UMG evaluiert wurde und dieses vorwiegend Themen aus dem Bereich der kardiopulmonalen Lehre umfasst, kann keine generelle Übertragbarkeit der gezogenen Schlüsse auf die gesamte Lehre im Studium der Humanmedizin erfolgen. Auch eine Generalisierbarkeit der Ergebnisse auf andere medizinische Fakultäten und insbesondere auf andere Studiengänge innerhalb und außerhalb Deutschlands kann nicht grundsätzlich vorausgesetzt werden. Daher sind weitere Studien zur Nutzung der Fremdeinschätzungs-Methode in Bezug auf spezifische Lernziele in anderen Modulen und Studiengängen notwendig.

Für die vorliegende Studie wurde ein zweizeitiges Studiendesign gewählt. Dieses ist jedoch zeit- und ressourcenaufwändiger als einzeitige Evaluationen. Zudem muss bei Erhebungen von Selbsteinschätzungen zu zwei Zeitpunkten der sogenannte *response shift bias* beachtet werden (siehe Kapitel 1.3.3). Ob dieser auch auf die Erhebung von Fremdeinschätzungen einen Einfluss hat bleibt unklar. Da in der vorliegenden Studie aufgrund der Notwendigkeit der zweizeitigen Datenerhebung der objektiven Prüfungsleistungen auch die Selbst- und

Fremdeinschätzungen zu zwei Zeitpunkten erhoben wurden, fand hier zu T2 keine Erhebung von rückblickenden Fremdeinschätzungen in Form von *thentests* statt. In weiterführenden Studien sollte somit die Möglichkeit einer einzeitigen Erhebung der Fremdeinschätzung durch sogenannte *thentests* geprüft werden.

Eine Stärke der vorliegenden Studie stellt wie zuvor erwähnt die Erhebung von objektiven Daten, zusätzlich zur Erhebung der Selbst- und Fremdeinschätzungen, dar.

4.4.2 Teilnahmerate und Prüfungskonsequenzen

Die Rücklaufquote betrug in der vorgestellten Studie 87,7 %. Diese ist vergleichbar mit den in der Literatur zur Vorhersage-basierten Evaluationsmethode beschriebenen Rücklaufquoten zwischen 60 – 94 % (Cohen-Schotanus et al. 2010; Schönrock-Adema et al. 2013).

Im Vorfeld der vorliegenden Studie wurde bereits in zwei vorangegangenen Studien (Wintersemester 2015/2016 und Wintersemester 2016/2017) der Versuch unternommen, den Einsatz der Fremdeinschätzungs-Methode zur Berechnung des studentischen Lernerfolgs zu prüfen. In einer ersten Studie wurden zur Erhebung der Fremdeinschätzungen keine Likert-Skalen verwendet, um eine größere Genauigkeit der Einschätzungen zu ermöglichen. Im Gegensatz zu den Fremdeinschätzungen wurden die Selbsteinschätzungen auf einer Likert-Skala erhoben. Dadurch ergab sich jedoch eine fehlende Kongruenz zwischen den Fremdeinschätzungen und den Selbsteinschätzungen und die Vergleichbarkeit beider Methoden war eingeschränkt. Zudem wurden in diesem ersten Versuch keine Leistungspunkte in der objektiven Prüfung zu T2 vergeben. Im Vergleich zu früheren Studien, in denen die gleiche Prüfung verwendet wurde, war die Erfolgsquote der objektiven Prüfung zum Zeitpunkt T2 jedoch deutlich geringer. Es wurde somit die Vermutung angestellt, dass die Studierenden aufgrund eines fehlenden Anreizes die objektive Prüfung nicht ernsthaft bearbeitet haben. In einem zweiten Versuch im Wintersemester 2016/2017 wurden nun zum einen die Selbst- und Fremdeinschätzungen auf einer kongruenten Likert-Skala erhoben und zudem Leistungspunkte in der objektiven Prüfung zu T2 vergeben. In diesem Durchlauf konnte jedoch im Nachhinein ein weiteres Problem identifiziert werden: Durch die gegebene Relevanz für die Studierenden zum Zeitpunkt T2 waren, vermutlich durch die Entwendung eines der papierbasierten Exemplare der objektiven Prüfung zu T1, die Aufgabenstellungen der objektiven Prüfung bekannt geworden. Hierdurch erzielten die Studierenden in der objektiven Prüfung zu T2 deutlich mehr Punkte als in vorherigen Studien. Somit stellte diese Prüfung kein repräsentatives Ergebnis des tatsächlichen Lernzuwachses der Studierenden dar. Aufgrund der Erkenntnisse aus den zwei Vorstudien wurde in der vorliegenden Studie zum einen eine kongruente Likert-Skala zur Erhebung der Selbst- und Fremdeinschätzungen genutzt und zudem eine neue, objektive und computerbasierte Prüfung konzipiert, die den Studierenden im Vorhinein nicht bekannt war und in der zu T2, zur Erhöhung der Motivation zur ernsthaften Beantwortung der Aufgaben, Leistungspunkte für das Modul 3.1 im Sinne einer summativen Prüfung vergeben wurden.

Aufgrund der Tatsache, dass es sich um einen möglichen Erwerb von Leistungspunkten handelte, kann davon ausgegangen werden, dass die Studierenden aktiv für die Prüfung zu T2 gelernt und sich intensiv mit den Themen befasst haben. Durch das aktive Lernen und Befassen mit der Thematik fiel es den Studierenden möglicherweise leichter, auch exaktere Selbst- und Fremdeinschätzungen abzugeben, da sie sich sowohl selbst aktiv mit dem Lernstoff auseinandergesetzt haben als auch sehr wahrscheinlich mit ihren Kommilitonen über die bevorstehende Prüfung austauschten. Dieser Austausch ist ein möglicher Grund für die gute Fähigkeit, Vorhersagen über den Leistungsstand der Kommilitonen zu treffen. Diese Vorhersagen waren jedoch den Selbsteinschätzungen nicht überlegen, sondern gleichwertig. Es bleibt insgesamt unklar, wie gut sich das Instrument für Module und Lehrveranstaltungen eignet, in denen es keine Prüfung und damit auch einen geringeren Anreiz für die Studierenden gibt, sich mit der Thematik zu beschäftigen.

Eine Stärke der vorgestellten Fremdeinschätzungs-Methode besteht darin, dass es nicht notwendig ist, dass die gleichen Studierenden an beiden Zeitpunkten an den Datenerhebungen teilnehmen. Zudem generiert die Fremdeinschätzungs-Methode, genau wie die Selbsteinschätzungs-Methode, Lernerfolgsdaten der Studierenden, ohne dass für die Abschätzung des Lernerfolgs objektive Prüfungen notwendig sind (Anders et al. 2016).

4.4.3 Lernziele

Wie in der Einleitung bereits angesprochen, kann durch die Ausrichtung von Evaluationsinstrumenten an Lernzielen, bei Nicht-Erreichen der vorgegebenen Lernziele, auf Defizite innerhalb des Curriculums geschlossen werden (Harden 1999). Die Ausrichtung an spezifischen Lernzielen war daher auch ein zentraler Aspekt der vorliegenden Studie. Die Erhebungen der Selbst- und Fremdeinschätzungen sowie die objektiven Prüfungen wurden an den Lernzielen aus dem Göttinger Lernzielkatalog ausgerichtet. Dabei wurde bereits in Kapitel 2.3.5 definiert, dass ein optimaler Lernzuwachs einem oAPG $\geq$ 50 % entspricht. In der vorliegenden Studie wiesen sechs von 27 Lernzielen ein oAPG $\leq$ 50 % auf, entsprechend einem suboptimalem Lernzuwachs. Die Lernziele behandelten dabei die folgenden Themen: „Chronische Bronchitis“, „Atherosklerose“, „Tuberkulose“, „Kardiomyopathie“, „Lungenfibrose“ sowie „Hyperlipidämie“. Es kann somit gefolgert werden, dass für diese sechs Lernziele ein Verbesserungsbedarf in der Lehre im Modul 3.1 besteht, da der objektiv gemessene Lernzuwachs unterhalb des gesetzten Grenzwertes eines optimalen Lernzuwachses lag. Es kann zudem festgestellt werden, dass sich unter diesen sechs Lernzielen sowohl kardiologische als auch pulmonologische Lernziele befanden und somit vermutlich kein allgemeines Defizit in der Lehre einer einzelnen Fachrichtung vorlag.

Wie in Kapitel 2.3.6 und 3.2 ausführlich beschrieben, fielen zu Beginn der Auswertung der Ergebnisse fünf Lernziele auf, die eine starke Abweichung zwischen dem sAPG und dem oAPG sowie dem pAPG und dem oAPG aufwiesen. Für einige der Lernziele konnten Erklärungen für die Abweichungen identifiziert werden, die in Kapitel 3.2 dargestellt sind.

Die Frage zum Lernziel „Perikarditis“ wurde nach ausführlicher Analyse aus der Wertung genommen, da uneindeutige Lehrinhalte möglicherweise zu einer Verwirrung der Studierenden geführt haben, wodurch diese nicht sicher die richtigen Antwortmöglichkeiten auswählen konnten. Es konnte im Nachhinein bei der Auswertung der Daten nicht sicher nachvollzogen werden, welche Lehrinhalte den Studierenden während der Vorlesung präsentiert wurden. Zudem ließen die zur Verfügung stehenden Lehrmaterialien uneindeutige Schlüsse hierüber zu. Auf dieser Basis wurde entschieden, dass eine Nicht-Berücksichtigung dieses Items die einzige Möglichkeit darstellte, eine Verfälschung der Daten durch ein möglicherweise fehlerhaftes Item auszuschließen.

Das Lernziel „Koronare Herzkrankheit“ zeigte eine fehlende Kongruenz zwischen den Selbst- und Fremdeinschätzungsaussagen und der Prüfungsfrage in der objektiven Prüfung zu ebendiesem Lernziel. Eine exakte Passung der Wortwahl ist jedoch eine Voraussetzung für die Inhaltsvalidität des Evaluationsinstrumentes mittels CSA (Anders et al. 2016). In der Sensitivitätsanalyse zeigte sich, dass die Selbst- und Fremdeinschätzungsaussagen allgemeiner formuliert waren als die Aufgabenstellung der objektiven Prüfung. Für Studierende sind Selbsteinschätzungen in Bezug auf ihren Leistungsstand für Lernziele mit sehr allgemein gehaltenen Definitionen möglicherweise schwierig zu bestimmen (Anders et al. 2016). Möglicherweise überschätzten die Studierenden durch eine fehlende exakte Wortwahl deshalb ihre eigenen Leistungen (Anders et al. 2016). Aus ebendiesem Grund der fehlenden Kongruenz und der damit einhergehenden fehlenden Inhaltsvalidität wurde entschieden, dieses Item aus der Wertung zu nehmen und in der nachfolgenden Analyse nicht weiter zu berücksichtigen.

Da sich für das Lernziel „Kardiomyopathie“ weder ein Fehler in der Item-Konstruktion, noch eine mangelnde Inhaltsvalidität nachweisen ließ, wurde entschieden das Item unverändert in der Auswertung zu belassen. Aufgrund der starken Abweichung zwischen dem sAPG/pAPG und oAPG von > 50 % für beide Methoden (Differenz zwischen sAPG und oAPG = 58,8 %; Differenz zwischen pAPG und oAPG = 61,6 %), hat dieses abweichende Lernziel jedoch einen starken Einfluss auf den Zusammenhang zwischen dem Lernzuwachs aus beiden Evaluationsmethoden und dem Lernzuwachs der objektiven Prüfung und vermindert folglich die Pearson-Korrelation r für den Gesamtzusammenhang des Lernzuwachses beider Evaluationsmethoden mit der objektiven Prüfung. Im Anhang A3 ist zu erkennen, dass die Itemschwierigkeiten der fünf Unteritems zu diesem Lernziel zu T1 zwischen 0,43 und 0,88 lagen. Zum Zeitpunkt T2 betrugen die Itemschwierigkeiten der ersten vier Unteritems Werte zwischen 0,79 und 0,86. Lediglich ein Unteritem zeigte eine hohe Itemschwierigkeit mit einem Wert von 0,18 und einer negativen Itemtrennschärfe von -0,01. Da jedoch kein erkennbarer Grund für die großen Abweichungen im Lernzuwachs zu finden war, muss davon ausgegangen werden, dass die Studierenden für das Lernziel „Kardiomyopathie“ nicht in der Lage waren, den objektiven Lernzuwachs durch die Angabe von Selbst- oder Fremdeinschätzungen adäquat einzuschätzen. Der Grund, weshalb diese Einschätzung für ebendieses Lernziel nicht möglich war, bleibt unklar.

Für die zwei Fragen zu den Lernzielen „Herzinsuffizienz“ und „Fallot’sche Tetralogie“ fiel aufgrund von Unstimmigkeiten zwischen der durchgeführten Lehre und der als korrekt gewerteten Antwortmöglichkeit jeweils eines Unteritems der Entschluss, dieses Unteritem in beiden Fällen umzukodieren. Es konnte in der Sensitivitätsanalyse nachgewiesen werden, dass die in der primären Analyse als „korrekt“ bewertete Antwortmöglichkeit in der stattgefundenen Lehre widersprüchlich gelehrt wurde. Diese Feststellung weist darauf hin, dass in der Lehre möglicherweise Inhalte vermittelt wurden, die nicht immer dem aktuellen Stand der Forschung entsprechen oder aber widersprüchliche Meinungen verschiedener Lehrender repräsentieren. Eine Konsequenz, die daraus gezogen werden kann, ist, dass Prüfungs- und Evaluationsinhalte sich nicht nur am aktuellen wissenschaftlichen Stand orientieren müssen, sondern zudem die durchgeführten Lehrinhalte bei der Erstellung dieser Beachtung finden sollten. Hierdurch können möglicherweise fehlende Kongruenzen aufgedeckt und vermieden sowie korrigiert werden und zudem eine hohe Lehrqualität gesichert werden.

Die beschriebenen Erkenntnisse demonstrieren eine Limitation der beiden Evaluationsmethoden. Zwar können die Evaluationen auf Mängel in der Lehre hindeuten, wenn Selbsteinschätzungen/Fremdeinschätzungen einen niedrigen Lernzuwachs anzeigen, jedoch zeigte sich in der vorliegenden Studie, dass auch im Falle eines hohen selbsteingeschätzten/vorhergesagten Lernzuwachses für einige Lernziele gleichzeitig ein niedriger objektiver Lernzuwachs vorliegen kann. Bezogen auf einige andere Lernziele in der vorliegenden Studie lag wiederum der objektiv gemessene Lernzuwachs höher als der selbsteingeschätzte/vorhergesagte Lernzuwachs zu ebendiesem Lernziel. In der vorliegenden Studie wurden diese Unterschiede aufgrund dessen detektiert, dass zusätzlich zur Erhebung der Selbst- und Fremdeinschätzungen eine objektive Prüfung durchgeführt wurde. Wenn ein Evaluationsinstrument jedoch regulär zur Evaluation der Lehre eingesetzt wird, erhalten die Evaluierenden in der Regel keine zusätzliche Auskunft über den tatsächlichen, objektiven Lernzuwachs, da dieser durch die Evaluationsinstrumente geeignet abgebildet wird. Allerdings muss hierbei stets bedacht werden, dass dieser in Einzelfällen abweichen kann, wenn beispielsweise die Kongruenz zwischen den Evaluationsaussagen und der objektiven Prüfung nicht ausreichend ist. Eine Schlussfolgerung, die hieraus gezogen werden kann, ist somit, dass für die Inhaltsvalidität der Evaluationsinstrumente mittels vergleichender studentischer Selbsteinschätzungen und Fremdeinschätzungen auf die exakte Passung der zu evaluierenden Lernziele und die Evaluationsaussagen geachtet werden muss (Raupach und Schiekirka – User manual for the CSA gain evaluation tool).

Bei Betrachtung desjenigen Lernziels mit dem niedrigsten Wert zum Zeitpunkt T1 in den Selbst- sowie Fremdeinschätzungen (entsprechend dem größten Vorwissen laut Selbst- bzw. Fremdeinschätzungen) fällt auf, dass dieses übereinstimmend das Item zum Lernziel „Atherosklerose“ ist. Diese Beobachtung bestätigt somit vor dem Hintergrund, dass die Thematik zur Atherosklerose im vorangehenden Studienjahr ausführlich gelehrt wird, die Validität der beiden Erhebungsmethoden.

4.5 Schlussfolgerungen für Evaluierende

In den folgenden Abschnitten sollen konkrete Schlussfolgerungen beschrieben werden, die aus den zuvor beschriebenen Ergebnissen gezogen werden können und den Evaluierenden bei der Interpretation ihrer Evaluationsergebnisse helfen sollen.

4.5.1 Rücklaufquoten

Um aus den ermittelten Daten der Rücklaufquoten einen konkreten Nutzen ziehen zu können, müssen diese in einem praktischen Kontext betrachtet werden. Wie zuvor beschrieben, lagen die notwendigen Rücklaufquoten für die Selbst- und Fremdeinschätzungs-Methode in der vorliegenden Studie bei 19,0 % bzw. 19,9 %. Diese beiden Methoden sind somit in Bezug auf den minimal notwendigen Evaluationsrücklauf gleichwertig. Das bedeutet konkret für die Evaluierenden, dass sie ab einem Rücklauf von knapp 20 % der Kohorte verlässliche Daten zum Lernzuwachs durch beide Methoden erhalten können. Allerdings ist, wie zuvor beschrieben, diese Quote abhängig vom tatsächlichen, objektiven Lernzuwachs. Es kann also nicht generell davon ausgegangen werden, dass eine Rücklaufquote von 20 % ausreichend ist, um verlässliche Werte zum Lernzuwachs zu erhalten. Jedoch kann gefolgert werden, dass bei einer Rücklaufquote von über 20 % der Kohorte für beide Methoden bei einem hinreichend hohen Lernzuwachs aus den Selbst- bzw. Fremdeinschätzungen, der tatsächliche Lernzuwachs in Bezug auf das Lernziel hoch war und die Evaluierenden den Ergebnissen vertrauen können. Für Lernziele, die in den Selbst- bzw. Fremdeinschätzungen einen niedrigen Lernzuwachs aufweisen, sollte die Aussagekraft der Evaluationen vorsichtig betrachtet werden, da die Rücklaufquote von 20 % der Kohorte möglicherweise nicht ausreichend ist, um reliable Ergebnisse zu erhalten. Eine mögliche Erklärung für die höhere NNE bei Lernzielen mit einem niedrigen objektiven Lernzuwachs wäre eine größere Streuung in diesem Bereich. Diese lässt sich beispielsweise dadurch erklären, dass der Leistungsunterschied von Studierenden innerhalb einer Kohorte stärker bei denjenigen Lernzielen zum Tragen kommt, die in der Lehre „schlecht“ gelehrt wurden und somit einen niedrigen oAPG aufweisen. In diesen Fällen ist die Streuung des objektiven Lernzuwachses der einzelnen Studierenden möglicherweise deshalb sehr viel größer als für Lernziele mit einem hohen oAPG, da trotz der „schlechten“ Lehre einige Studierende eigenständig viel lernen und dementsprechend trotz allem einen hohen Lernzuwachs für diese Lernziele erzielen. Diejenigen Studierenden, die sich lediglich auf die durchgeführte Lehre verlassen und nicht zusätzlich eigenständig lernen, würden in diesen Fällen vermutlich einen deutlich niedrigeren Lernzuwachs bezüglich der genannten Lernziele aufweisen. Um aus einer Kohorte mit großer Streuung eine NNE zu erhalten welche reliable Daten für die Gesamtkohorte liefert, ist somit eine deutlich höhere Anzahl an Studierenden notwendig, um ebendieses Ergebnis der Gesamtkohorte korrekt widerzuspiegeln. Eine weitere denkbare Erklärung wäre eine Unsicherheit von Studierenden bezüglich der Lernziele, die nur zu einem kleinen Lernzuwachs geführt haben. Dieser Zusammenhang sollte in

zukünftigen Studien weiter untersucht werden. Ob eine unterschiedlich große Stichprobe, im Sinne einer größeren oder kleineren Studienkohorte als in der vorliegenden Studie, einen Einfluss auf die Höhe der notwendigen Rücklaufquote hat, kann hier nicht abschließend geklärt werden und sollte in weiterführenden Studien überprüft werden.

Die Konsequenz dieser Ergebnisse ist die größere Flexibilität in der Durchführung der lernzielbasierten Lernerfolgsevaluation, da sowohl die Selbsteinschätzungs-Methode als auch die Fremdeinschätzungs-Methode niedrige Rücklaufquoten erfordern und gleichzeitig eine hohe Korrelation mit dem objektiv gemessenen Lernzuwachs sowie untereinander aufweisen. Es kann somit gefolgert werden, dass in diesem Kontext beide Methoden gleichwertig verwendbar sind.

4.5.2 Differenzierung zwischen gut und schlecht gelehrten Lernzielen

Um eine praktische Nutzung und Interpretation im Evaluationsalltag zu erleichtern, wurde in der vorgestellten Studie eine ROC-Analyse durchgeführt, die der Bestimmung eines *Cut-off*-Wertes zur Unterscheidung zwischen optimalem und suboptimalem Lernzuwachs diente (siehe Kapitel 3.6). Hierbei ergaben sich für beide Evaluationsmethoden optimale *Cut-off*-Werte von 61 %, das heißt bei Werten darüber kann von einem optimalen, darunter von einem suboptimalen Lernzuwachs gesprochen werden. Für die Fremdeinschätzungs-Methode lag bei diesem *Cut-off*-Wert eine Sensitivität von 90 % vor, was für den praktischen Einsatz bedeutet, dass für 90 % der Lernziele, die in der objektiven Prüfung einen optimalen Lernzuwachs aufwiesen, auch in der Vorhersage-basierten Evaluationsmethode mittels Fremdeinschätzungen ein optimaler Lernzuwachs erzielt wurde. Der PPW von 95 % beschreibt dabei, dass lediglich 5 % der durch die Fremdeinschätzungen als gut gelehrte identifizierten Lernziele in der objektiven Prüfung keinen optimalen Lernzuwachs aufwiesen, während 95 % der durch die Fremdeinschätzungen als optimal gelehrt identifizierten Lernziele auch objektiv einen guten Lernzuwachs zeigten. Für die Selbsteinschätzungs-Methode ergab der *Cut-off*-Wert von 61 % eine Sensitivität von 86 %, entsprechend einer hohen Rate von Lernzielen, die durch die Selbsteinschätzungs-Methode ebenfalls einen hohen Lernzuwachs aufwiesen, wenn sie diesen in der objektiven Prüfung erzielten. Gleichzeitig betrug der PPW für die Selbsteinschätzungs-Methode, genau wie für die Fremdeinschätzungs-Methode, 95 %. Lehrende können hieraus bei der Interpretation der Evaluationsergebnisse folgende Schlüsse ziehen: Für Lernziele, die in Evaluationen durch die Selbst- oder die Fremdeinschätzungs-Methode einen Lernzuwachs von mindestens 61 % aufweisen, besteht eine sehr große Wahrscheinlichkeit, dass diese auch einen hohen (entsprechend einem optimalen) objektiven Lernzuwachs aufweisen. Diese Lernziele wurden somit mit großer Wahrscheinlichkeit ausreichend gut durch die stattgefundene Lehre gelehrt, sodass das vorgegebene Ziel, einen Lernzuwachs von mindestens 50 % zu erzielen, für diese Lernziele erreicht wurde. In Bezug auf diese Lernziele müssen sich Lehrende somit keine Sorgen hinsichtlich einer unzureichenden stattgefundenen Lehre machen.

Bei diesem *Cut-off*-Wert von 61 % und einer gleichzeitig vorliegenden Spezifität von 83 % für beide Methoden wird zudem ein Großteil der tatsächlich, objektiv nicht vollständig erreichten Lernziele durch die Selbst- bzw. Fremdeinschätzungs-Methode als solche identifiziert. Der NPW betrug hierbei 71 % für die Fremdeinschätzungs-Methode, sodass ebendieser Anteil der durch die Fremdeinschätzungen als suboptimal gelehrt identifizierten Lernziele auch objektiv einen suboptimalen Lernzuwachs aufwiesen. Für die Selbsteinschätzungs-Methode betrug der NPW 62 %, sodass mehr als die Hälfte der Lernziele, die durch die Selbsteinschätzungen einen suboptimalen Lernzuwachs aufwiesen, diesen auch in der objektiven Prüfung zeigten. Es kann hieraus geschlussfolgert werden, dass Lernziele, die in den Selbst- oder Fremdeinschätzungen einen suboptimalen Lernzuwachs (< 61 %) aufweisen, von den Lehrenden genauer betrachtet und die stattgefundene Lehre hierzu überprüft werden sollte, da ein Großteil der als suboptimal identifizierten Lernziele auch tatsächlichen einen suboptimalen, objektiven Lernzuwachs aufweist. Diesbezüglich wurden somit die festgesetzten Ziele bezogen auf den Lernzuwachs nicht erreicht, weshalb die Lehre hierzu kritisch hinterfragt werden sollte. In diesen Fällen sollten sich die Lehrenden mit der Frage beschäftigen, ob und wodurch die Lehre (in Bezug auf diese spezifischen Lernziele) dahingehend verbessert werden kann, dass Studierende die vorgegebenen Lernziele erreichen. Darüber hinaus besteht bei einer Spezifität von 83 % für beide Methoden zudem ein gewisses Risiko, dass suboptimal gelehrte Lernziele nicht als solche erkannt werden. Aus diesem Grund sollten Lehrende sich unabhängig vom Evaluationsergebnis stets kritisch hinterfragen. Sie sollten zusätzlich zur schriftlichen Evaluation mittels der Erhebung von Selbst- oder Fremdeinschätzungen, die Rückmeldungen der Studierenden, beispielsweise in Form von Freitextkommentaren in der Evaluation, zur Kenntnis nehmen und die Ergebnisse als Ergänzung zur Verbesserung der Lehre ansehen.

Die Ergebnisse der ROC-Analyse erlaubten somit die Festlegung eines *Cut-off*-Wertes, der für beide Methoden eine weitestgehend verlässliche Identifizierung von optimalem bzw. suboptimalem Lernzuwachs für spezifische Lernziele liefert. Nichtsdestotrotz sollte weiterhin aufmerksam darauf geachtet werden, ob es Hinweise für Mängel oder Verbesserungsmöglichkeiten in Bezug auf die durchgeführte Lehre gibt, selbst wenn die Evaluationsergebnisse der Lernerfolgsevaluation mittels Selbst- oder Fremdeinschätzungen einen optimalen Lernzuwachs für spezifische Lernziele liefern.

4.6 Ausblick

Die vorliegende Studie trägt zur Weiterentwicklung der Lernerfolgsevaluation und zum Verständnis der Nutzung von Vorhersagen zur Evaluation der Lehre bei. Im Folgenden soll ein Ausblick auf diejenigen Fragen gegeben werden, welche die vorliegende Studie aufwarf und die Grundlage künftiger Forschungsprojekte sein könnten.

In folgenden Studien sollten weitere Bereiche im Medizinstudium und darüber hinaus einbezogen werden, deren Thematik und Lernziele zusätzliche thematische Bereiche der Lehre

abdecken. Wie Schiekirka et al. (2013) bereits für das Evaluationsinstrument mittels sAPG vorschlugen, sollte auch für das Instrument der Fremdeinschätzungs-Methode getestet werden, ob dieses eine Anwendungsmöglichkeit bezüglich der Evaluation praktischer Fertigkeiten und affektiver Lernziele bieten kann. Auch eine Nutzung in anderen Kontexten in der Hochschullehre ist denkbar, wenn dort eine Definition von spezifischen Lernzielen möglich ist. Hierzu müssten weiterführende Studien in anderen Studiengängen durchgeführt werden.

Anders et al. (2016) zeigten, dass die Nutzung sog. *qualifiers* für das Instrument der Lernerfolgsevaluation mittels sAPG zu einem leicht niedrigeren Lernerfolg durch die Selbsteinschätzungen führte, damit aber zugleich die Vergleichbarkeit des auf diese Weise gemessenen Lernerfolgs zwischen unterschiedlichen Lernzielen erhöhte. Diese *qualifiers*, die zu einer Spezifizierung der Selbsteinschätzungs-Aussagen führen, wurden in der vorliegenden Studie nicht aktiv eingesetzt. In Folgestudien könnte durch die Nutzung von *qualifiers* in der Item-Konstruktion der Einfluss dieser auf die Fremdeinschätzungs-Methode untersucht werden. Hierdurch kann möglicherweise eine höhere Korrelation und damit einhergehend eine bessere Spiegelung des tatsächlichen Lernerfolges durch den Lernerfolg aus Fremdeinschätzungen erreicht werden. Da durch die Nutzung der *qualifiers* der selbsteingeschätzte und vermutlich auch der fremdeingeschätzte Lernerfolg etwas niedriger ausfällt als ohne die Nutzung dieser, kann vermutet werden, dass durch deren Einsatz eine noch bessere Kongruenz zum objektiv gemessenen Lernzuwachs erzielt werden kann, da beide Evaluationsmethoden den objektiv gemessenen Lernerfolg in der vorliegenden Studie leicht überschätzten.

Zur Erleichterung der Durchführung der Evaluationen sollte in zukünftigen Studien zudem erforscht werden, ob eine einzeitige, retrospektive Erhebung des fremdeingeschätzten Leistungsstandes der Kommilitonen nach dem Modul ausreichend ist, um hieraus den Lernzuwachs der Studierenden zu berechnen. Wie bereits beschrieben, ist für die sAPG-Methode keine zweizeitige Datenerhebung notwendig (Schiekirka et al. 2014). Ob diese Beobachtung sich in gleicher Weise auf die Fremdeinschätzungs-Methode anwenden lässt, sollte Gegenstand zukünftiger Forschungsprojekte sein.

5 Zusammenfassung

Evaluationen stellen einen wichtigen Bestandteil eines jeden Lehrcurriculums dar, wobei in vielen Fällen studentische Bewertungen zur Evaluation der Lehre herangezogen werden. Sie erlauben es, Rückschlüsse über die Lehrqualität zu ziehen und auf Basis dieser zur Verbesserung der Lehre beizutragen. Speziell im Medizinstudium sollte dabei ein besonderer Fokus auf der Ergebnisqualität liegen. Die Nutzung von eindeutig definierten Lernzielen stellt eine wichtige Grundlage hierfür dar. In der Literatur ist vielfach die Nutzung studentischer Selbsteinschätzungen zur Evaluation beschrieben. Jedoch sind diese nicht unumstritten, was den Einfluss möglicher *bias* betrifft. Nichtsdestotrotz existieren valide Evaluationsinstrumente, die auf studentischen Selbsteinschätzungen beruhen. Ein Beispiel stellt die an der Universitätsmedizin Göttingen genutzte Lernerfolgsevaluation mittels vergleichender studentischer Selbsteinschätzungen dar. Zur Weiterentwicklung dieses Instrumentes wurde in der vorliegenden Studie „Erprobung der Vorhersage-basierten Evaluationsmethode für das Instrument der Lernerfolgsevaluation“ eine neue Herangehensweise, durch Nutzung der Vorhersage-basierten Evaluationsmethode in Kombination mit der Lernerfolgsevaluation, erprobt.

Im Rahmen dieser Forschungsarbeit wurden folgende Forschungsfragen beantwortet: Wie hängen die mittels der modifizierten Fremdeinschätzungs-Methode ermittelten Lernerfolgsdaten mit den Lernerfolgsdaten der ursprünglich genutzten Selbsteinschätzungs-Methode zusammen? Welcher Rücklauf muss mit beiden Methoden minimal erreicht werden, um reliable Lernerfolgsdaten zu erhalten? Wie hängt der aus Selbsteinschätzungen und Fremdeinschätzungen errechnete Lernerfolg mit dem in wiederholten Prüfungen erhobenen objektiven Lernerfolg zusammen?

Die Studie konnte eine hohe signifikante Korrelation zwischen dem Lernzuwachs der Evaluationsmethode mittels vergleichender studentischer Selbsteinschätzungen und dem Lernzuwachs der Fremdeinschätzungs-Methode nachweisen ($r = 0{,}952$; $p < 0{,}001$). Die Studierenden erzielten durch Selbsteinschätzungen im Mittel einen Lernzuwachs von $67{,}8 \pm 15{,}6$ % und durch Fremdeinschätzungen einen mittleren Lernzuwachs von $67{,}8 \pm 15{,}2$ % pro Lernziel. Beide Methoden konnten in der vorliegenden Studie durch die Nutzung der gleichen, sechsstufigen Likert-Skala exakt miteinander verglichen werden und zeigten in einem Bland-Altman-Plot eine sehr gute Vergleichbarkeit des Lernzuwachses (mittlere Differenz des Lernzuwachses pro Lernziel: 0,002 %). Diese gute Vergleichbarkeit konnte in der vorliegenden Studie jedoch nur für Lernziele mit einem Lernzuwachs von > 30 % bestätigt werden, da unterhalb dieser Grenze keine Daten zum Vergleich beider Methoden vorlagen. Bezüglich der notwendigen Rücklaufquoten beider Evaluationsmethoden zeigte sich ein vergleichsweise niedriger, minimal notwendiger Rücklauf von 19,0 % für die Selbsteinschätzungs-Methode und 19,9 % für die Fremdeinschätzungs-Methode, bezogen auf die Gesamtkohorte ($n = 128$). Der erforderliche Mindestrücklauf beider Methoden war damit vergleichbar, lies sich jedoch durch die Nutzung der Fremdeinschätzungs-

Methode im Gegensatz zur Selbsteinschätzungs-Methode nicht reduzieren. Eine weitere Erkenntnis konnte durch den Zusammenhang der notwendigen Rücklaufquoten und dem dazugehörigen objektiv gemessenen Lernzuwachs gewonnen werden. Hierbei zeigte sich, dass die notwendige Rücklaufquote für Lernziele mit einem hohen objektiven Lernzuwachs niedriger ausfiel als für Lernziele mit einem niedrigen objektiven Lernzuwachs. Diese Beobachtung konnte ebenfalls für beide Evaluationsmethoden gemacht werden. Zudem konnte eine hohe Korrelation zwischen dem Lernzuwachs aus Selbst- bzw. Fremdeinschätzungen und dem Lernzuwachs in einer objektiven Prüfung bestätigt werden (Zusammenhang des Lernzuwachses aus Selbsteinschätzungen und objektiven Prüfungsleistungen: $r = 0{,}709$; $p < 0{,}001$, Zusammenhang des Lernzuwachses aus Fremdeinschätzungen und objektiven Prüfungsleistungen: $r = 0{,}704$; $p < 0{,}001$). Insgesamt zeigten sich damit beide Methoden geeignet zur Nutzung als Evaluationsinstrument für die Lernerfolgsevaluation mittels spezifischer Lernziele.

Durch die Nutzung spezifischer Lernziele besaß die vorliegende Studie einige Vorteile gegenüber globalen Evaluationsmethoden. Des Weiteren wurde durch die Einbeziehung des initialen Leistungsstandes der tatsächlich durch die stattgefundene Lehre aufgetretene Lernzuwachs erhoben.

Die vorliegende Forschungsarbeit bestätigt somit die Validität der Evaluationsmethode mittels vergleichender studentischer Selbsteinschätzungen. Zudem unterstützt sie die Annahme, dass die Fremdeinschätzungs-Methode ein geeignetes Instrument zur Evaluation der Lehre sein kann. Diese zeigte sich in Bezug auf die Lernerfolgsevaluation gleichwertig, jedoch nicht überlegen gegenüber der Evaluationsmethode mittels vergleichender studentischer Selbsteinschätzungen.

6 Anhang

Tabelle A1: Itemanalyse der fünf abweichenden Lernziele auf Lernzielebene (entspricht: Fragenebene) für T1:

Item (Frage)	Minimum T1 Original-Punktzahl	Maximum T1 Original-Punktzahl	MW T1 Original-Punktzahl	Standardabweichung T1 Original-Punktzahl	Itemschwierigkeit T1	Itemtrennschärfe T1
7	0	5	2,46	1,12	0,49	0,07
24	0	5	2,33	1,02	0,47	0,13
25	0	5	3,30	1,07	0,66	0,09
28	1	5	2,79	0,95	0,56	0,14
29	0	5	2,38	1,22	0,48	0,00

Tabelle A2: Itemanalyse der fünf abweichenden Lernziele auf Lernzielebene (entspricht: Fragenebene) für T2:

Item (Frage)	Minimum T2 Original-Punktzahl	Maximum T2 Original-Punktzahl	MW T2 Original-Punktzahl	Standardabweichung T2 Original-Punktzahl	Itemschwierigkeit T2	Itemtrennschärfe T2
7	1	5	3,25	0,92	0,65	0,09
24	1	5	3,38	0,93	0,68	0,02
25	1	5	3,47	0,95	0,69	0,10
28	1	5	4,01	0,54	0,80	0,25
29	1	5	3,30	0,89	0,66	0,12

Tabelle A3: Itemanalyse der fünf abweichenden Lernziele auf Unteritem-Ebene für T1:

Item	Unter-item	Minimum T1 Original-Punktzahl	Maximum T1 Original-Punktzahl	MW T1 Original-Punktzahl	Stan-dardab-weichung T1 Original-Punktzahl	Itemschwie-rigkeit T1	Itemtrenn-schärfe T1
7	1	0	1	0,53	0,50	0,53	0,11
7	2	0	1	0,38	0,49	0,38	-0,01
7	3	0	1	0,72	0,45	0,72	-0,07
7	4	0	1	0,45	0,50	0,45	0,02
7	5	0	1	0,39	0,49	0,39	0,14
24	1	0	1	0,46	0,50	0,46	0,06
24	2	0	1	0,45	0,50	0,45	0,16
24	3	0	1	0,71	0,46	0,71	-0,10
24	4	0	1	0,49	0,50	0,49	0,01
24	5	0	1	0,21	0,41	0,21	0,05
25	1	0	1	0,54	0,50	0,54	0,07
25	2	0	1	0,65	0,48	0,65	0,09
25	3	0	1	0,43	0,50	0,43	0,04
25	4	0	1	0,88	0,32	0,88	0,10
25	5	0	1	0,80	0,40	0,80	0,04
28	1	0	1	0,20	0,40	0,20	-0,21
28	2	0	1	0,87	0,34	0,87	-0,07
28	3	0	1	0,66	0,48	0,66	0,26
28	4	0	1	0,40	0,49	0,40	0,02
28	5	0	1	0,67	0,47	0,67	0,16
29	1	0	1	0,41	0,49	0,41	-0,03
29	2	0	1	0,55	0,50	0,55	0,05
29	3	0	1	0,49	0,50	0,49	0,13
29	4	0	1	0,34	0,47	0,34	0,10
29	5	0	1	0,59	0,49	0,59	-0,02

Tabelle A4: Itemanalyse der fünf abweichenden Lernziele auf Unteritem-Ebene für T2:

Item	Un-ter-i-tem	Minimum T2	Maximum T2	MW T2 Original-Punktzahl	Stan-dardab-weichung T2 Origi-nal-Punktzahl	Itemschwie-rigkeit T2	Itemtrenn-schärfe T2
7	1	0	1	0,79	0,41	0,79	-0,12
7	2	0	1	0,45	0,50	0,45	-0,07
7	3	0	1	0,70	0,46	0,70	0,01
7	4	0	1	0,74	0,44	0,74	0,08
7	5	0	1	0,57	0,50	0,57	0,03
24	1	0	1	0,28	0,45	0,28	-0,05
24	2	0	1	0,82	0,39	0,82	0,15
24	3	0	1	0,96	0,19	0,96	0,14
24	4	0	1	0,97	0,17	0,97	0,23
24	5	0	1	0,34	0,48	0,34	0,05
25	1	0	1	0,84	0,37	0,84	0,17
25	2	0	1	0,79	0,41	0,79	0,08
25	3	0	1	0,86	0,35	0,86	0,20
25	4	0	1	0,80	0,40	0,80	0,01
25	5	0	1	0,18	0,39	0,18	-0,01
28	1	0	1	0,95	0,21	0,95	0,17
28	2	0	1	0,99	0,09	0,99	0,32
28	3	0	1	0,98	0,12	0,98	0,21
28	4	0	1	0,12	0,32	0,12	0,07
28	5	0	1	0,96	0,19	0,96	0,26
29	1	0	1	0,12	0,32	0,12	0,20
29	2	0	1	0,70	0,46	0,70	0,10
29	3	0	1	0,88	0,33	0,88	-0,06
29	4	0	1	0,77	0,42	0,77	-0,07
29	5	0	1	0,84	0,37	0,84	0,18

7 Literaturverzeichnis

Albanese MA, Kent TH, Whitney DR (1979): Cluing in multiple-choice test items with combinations of correct responses. J Med Educ 54, 948–950

Aleamoni LM (1999): Student rating myths versus research facts from 1924 to 1998. J Pers Eval Educ 13, 153–166

Altman DG, Bland JM (1983): Measurement in medicine: The analysis of method comparison studies. Statistician 32, 307–317

Anders S, Pyka K, Mueller T, von Streinbuechel N, Raupach T (2016): Influence of the wording of evaluation items on outcome-based evaluation results for large-group teaching in anatomy, biochemistry and legal medicine. Ann Anat 208, 222–227

Anziani H, Durham J, Moore U (2008): The relationship between formative and summative assessment of undergraduates in oral surgery. Eur J Dent Educ 12, 233–238

Armitage P, Berry G, Matthews JNS: Statistical Methods in Medical Research. 4. Auflage; Blackwell Science, Oxford 2002

Berger U, Schleußner C (2003): Hängen Ergebnisse einer Lehrveranstaltungs-Evaluation von der Häufigkeit des Veranstaltungsbesuches ab? Z Padagog Psycol 17, 125–131

Berk RA (2013): Top five flashpoints in the assessment of teaching effectiveness. Med Teach 35, 15–26

Billings-Gagliardi S, Barrett SV, Mazor KM (2004): Interpreting course evaluation results: insights from thinkaloud interviews with medical students. Med Educ 38, 1061–1070

Blanch-Hartigan D (2011): Medical students' self-assessment of performance: Results from three meta-analyses. Patient Educ Couns 84, 3–9

Bland JM, Altman DG (1986): Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1, 307–310

Bland JM, Altman DG (1997): Cronbach's alpha. BMJ 314, 572

Bland JM, Altman DG (1999): Measuring agreement in method comparison studies. Stat Methods Med Res 8, 135–160

Bloom BS, Engelhart MD, Furst EJ, Hill W, Krathwohl D: Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook 1: Cognitive Domain. David McKay Company, New York 1956

Braun E, Leidner B (2009): Academic course evaluation: Theoretical and empirical distinctions between self-rated gain in competences and satisfaction with teaching behavior. Eur Psychol 14, 297–306

Centra JA (2003): Will teachers receive higher student evaluations by giving higher grades and less course work? Res High Educ 44, 495-518

Cohen J: Statistical Power Analysis for the Behavioral Sciences. 2. Auflage; Lawrence Erlbaum Associates, Hillsdale, N.J. 1988

Cohen-Schotanus J, Schönrock-Adema J, Schmidt HG (2010): Quality of courses evaluated by 'predictions' rather than opinions: Fewer respondents needed for similar results. Med Teach 32, 851–856

Colthart I, Bagnall G, Evans A, Allbutt H, Haig A, Illing J, McKinstry B (2008): The effectiveness of self-assessment on the identification of learner needs, learner activity, and impact on clinical practice: BEME Guide No. 10. Med Teach 30, 124–145

Cummings SM, Savitz LA, Konrad TR (2001): Reported response rates to mailed physician questionnaires. Health Serv Res 35, 1347–1355

d'Apollonia S, Abrami PC (1997): Navigating student ratings of instruction. Am Psychol 52, 1198–1208

Davis DA, Mazmanian PE, Fordis M, van Harrison R, Thorpe KE, Perrier L (2006): Accuracy of physician self-assessment compared with observed measures of competence: A systematic review. JAMA 296, 1094–1102

D'Eon M, Sadownik L, Harrison A, Nation J (2008): Using self-assessments to detect workshop success. Am J Eval 29, 92–98

Drennan J, Hyde A (2008): Controlling response shift bias: The use of the retrospective pretest design in the evaluation of a master's programme. Assess Eval High Educ 33, 699–709

Dybowski C, Sehner S, Harendza S (2017): Influence of motivation, self-efficacy and situational factors on the teaching quality of clinical educators. BMC Med Educ 17, 84

Epstein RM, Hundert EM (2002): Defining and assessing professional competence. JAMA 287, 226–235

Eva KW, Regehr G (2005): Self-assessment in the health professions: A reformulation and research agenda. Acad Med 80, 46-54

Fitzgerald JT, White CB, Gruppen LD (2003): A longitudinal study of self-assessment accuracy. Med Educ 37, 645–649

Frank JR, Danoff D (2007): The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Med Teach 29, 642–647

Franklin J, Theall M: Who reads ratings: Knowledge, attitude, and practice of users of student ratings of instruction: Vortrag im Rahmen des Annual Meeting of the American Educational Research Association, San Francisco, 27.03.–31.03.1989. ERIC ED 306 241. Zitiert nach Inhaltsangabe des Vortrages (gehalten 1989)

Germain ML, Scandura TA (2005): Grade inflation and student individual differences as systematic bias in faculty evaluations. J Instruc Psychol 32, 58–67

Gibson KA, Boyle P, Black DA, Cunningham M, Grimm MC, McNeil HP (2008): Enhancing evaluation in an undergraduate medical education program. Acad Med 83, 787–793

Gliem JA, Gliem RR: Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales: Vortrag im Rahmen der Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, Columbus, 08.10–10.10.2003. Zitiert nach Inhaltsangabe des Vortrages (gehalten 2003)

Gordon MJ (1991): A review of the validity and accuracy of self-assessments in health professions training. Acad Med 66, 762–769

Greenwald AG (1997): Validity concerns and usefulness of student ratings of instruction. Am Psychol 52, 1182–1186

Greenwald AG, Gillmore GM (1997): Grading leniency is a removable contaminant of student ratings. Am Psychol 52, 1209–1217

Griffin BW (2001): Instructor reputation and student ratings of instruction. Contemp Educ Psychol 26, 534–552

Güntert A, Wanner E, Brauer HP, Stobrawa FF: Approbationsordnung für Ärzte (ÄAppO) Bundesärzteordnung (BÄO): Mit Erläuterungen und praktischen Hinweisen. Deutscher Ärzte-Verlag, Köln 2003

Harden RM (1999): AMEE Guide No. 14: Outcome-based education: Part 1 – An introduction to outcome-based education. Med Teach 21, 7–14

Harden RM (2007a): Learning outcomes as a tool to assess progression. Med Teach 29, 678–682

Harden RM (2007b): Outcome-based education – the ostrich, the peacock and the beaver. Med Teach 29, 666–671

Herzig S, B. Marschall D, Nast-Kolb S, Soboll LC, Rump and R. D. Hilgers (2007): Positionspapier der nordrhein-westfälischen Studiendekane zur hochschulvergleichenden leistungsorientierten Mittelvergabe für die Lehre. GMS Z Med Ausbild 24, Doc. 109

Hofstee WKB, Schaapmann H (1990): Bets beat polls: Averaged predictions of election outcomes. Acta Politica 25, 257–270

Honest H, Khan KS (2002): Reporting of measures of accuracy in systematic reviews of diagnostic literature. BMC Health Serv Res 2, 4

Howard GS (1980): Response-shift bias – A problem in evaluating interventions with pre/post self-reports. Eval Rev 4, 93–106

Howard GS, Maxwell SE (1980): Correlation between student satisfaction and grades: A sase of mistaken causation? J Educ Psychol 72, 810–820

Howard GS, Ralph KM, Gulanick NA, Maxwell SE, Nance DW, Gerber SK (1979): Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. Appl Psychol Meas 3, 1–23

Kanzow P, Schuelper N, Witt D, Wassmann T, Sennhenn-Kirchner S, Wiegand A, Raupach T (2018): Effect of different scoring approaches upon credit assignment when using multiple true-false items in dental undergraduate examinations. Eur J Dent Educ 22, e669-e678

Kellerman S (2001): Physician response to surveys: A review of the literature. Am J Prev Med 20, 61–67

Kibble JD (2017): Best practices in summative assessment. Adv Physiol Educ 41, 110–119

Kogan JR, Shea JA (2007): Course evaluation in medical education. Teach Teach Educ 23, 251–264

Kruger J, Dunning D (1999): Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. J Pers Soc Psychol 77, 1121–1134

Lam TCM (2009): Do self-assessments work to detect workshop success? An analysis of argument and recommendation by D'Eon et al. Am J Eval 30, 93–105

Leopold SS, Morgan HD, Kadel NJ, Gardner GC, Schaad DC, Wolf FM (2005): Impact of educational intervention on confidence and competence in the performance of a simple surgical task. J Bone Joint Surg Am 87, 1031–1037

Linse AR (2017): Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. Stud Educ Eval 54, 94–106

Manthei RJ (1997): The response-shift bias in a counsellor education programme. Br J Guid Counc 25, 229–237

Marsh HW (1984): Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. J Educ Psychol 76, 707–754

Marsh HW (1987): Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. Int J Educ Res 11, 253–388

Marsh HW, Roche LA (1997): Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. Am Psychol 52, 1187–1197

Mason KH, Edwards RR, Roach DW (2002): Student evaluation of instructors: A measure of teaching effectiveness or of something else? Journal of Business Administration Online 1, 1–11

McKeachie WJ (1997): Student ratings: The validity of use. Am Psychol 52, 1218–1225

McOwen KS, Bellini LM, Morrison G, Shea JA (2009): The development and implementation of a health-system-wide evaluation system for education activities: Build it and they will come. Acad Med 84, 1352–1359

Metz CE (1978): Basic principles of ROC analysis. Semin Nucl Med 8, 283–298

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

Möltner A, Duelli R, Resch F, Schultz J-H, Jünger J (2010): Fakultätsinterne Prüfungen an den deutschen medizinischen Fakultäten. GMS Z Med Ausbild 27, Doc44

Möltner A, Schellberg D, Jünger J (2006): Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild 23, Doc53

Müller T, Montano D, Poinstingl H, Dreiling K, Schiekirka-Schwake S, Anders S, Raupach T, Steinbüchel N von (2017): Evaluation of large-group lectures in medicine – development of the SETMED-L (Student Evaluation of Teaching in MEDical Lectures) questionnaire. BMC Med Educ 17, 137

Murray HG: Student evaluation of teaching: Has it made a difference?: Vortrag im Rahmen des Annual Meeting of the Society for Teaching and Learning in Higher Education, Charlottetown, Juni 2005. Zitiert nach Inhaltsangabe des Vortrages (gehalten 2005)

NKLM: siehe MFT Medizinischer Fakultätentag der Bundesrepublik Deutschland e. V. 2015

Norman G (2003): Hi! How are you? Response shift, implicit theories and differing epistemologies. Qual Life Res 12, 239–249

O'Rourke T (1999): Increasing response rates: Specific applications. Am J Health Stud 15, 164–166

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

Raupach T, Harendza S, Anders S, Schuelper N, Brown J (2016): How can we improve teaching of ECG interpretation skills? Findings from a prospective randomised trial. J Electrocardiol 49, 7–12

Raupach T, Münscher C, Beissbarth T, Burckhardt G, Pukrop T (2011): Towards outcome-based programme evaluation: Using student comparative self-assessments to determine teaching effectiveness. Med Teach 33, e446-453

Raupach T, Schiekirka S, Münscher C, Beißbarth T, Himmel W, Burckhardt G, Pukrop T (2012): Piloting an outcome-based programme evaluation tool in undergraduate medical education. GMS Z Med Ausbild 29, Doc44

Rindermann H (2003): Lehrevaluation an Hochschulen: Schlussfolgerungen aus Forschung und Anwendung für Hochschulunterricht und seine Evaluation. Zeitschrift für Evaluation 3, 233–256

Rindermann H, Schofield N (2001): Generalizability of multidimensional student ratings of university instruction across courses and teachers. Res High Educ 42, 377–399

Roediger HL, Karpicke JD (2006): The power of testing memory: Basic research and implications for educational practice. Perspect Psychol Sci 1, 181–210

Rohs FR (1999): Response shift bias: A problem in evaluating leadership development with self-report pretest-posttest measures. JAE 40, 28–37

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

Schiekirka S, Feufel MA, Herrmann-Lingen C, Raupach T (2015): Evaluation in medical education: A topical review of target parameters, data collection tools and confounding factors. Ger Med Sci 13, Doc15

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

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

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

Schiekirka-Schwake S, Barth J, Pfeilschifter J, Hickel R, Raupach T, Herrmann-Lingen C (2019): National survey of evaluation practices and performance-guided resource allocation at German medical schools. Ger Med Sci 17, Doc04

Schönrock-Adema J, Lubarsky S, Chalk C, Steinert Y, Cohen-Schotanus J (2013): 'What would my classmates say?' An international study of the prediction-based method of course evaluation. Med Educ 47, 453–462

Seldin P (1989): Using student feedback to improve teaching. New Dir Teach Learn 37, 89–97

Sennhenn-Kirchner S, Goerlich Y, Kirchner B, Notbohm M, Schiekirka S, Simmenroth A, Raupach T (2018): The effect of repeated testing vs repeated practice on skills learning in undergraduate dental education. Eur J Dent Educ 22, e42-e47

Shumway JM, Harden RM (2003): AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. Med Teach 25, 569–584

Simmenroth A, Chenot JF, Fischer T, Scherer M, Stanske B, Kochen M (2007): Mit Laienschauspielern das ärztliche Gespräch trainieren. Dtsch Arztebl 104, 851–852

Skeff KM, Stratos GA, Bergen MR (1992): Evaluation of a medical faculty development program: A comparison of traditional pre/post and retrospective pre/post self-assessment ratings. Eval Health Prof 15, 350–366

Spady WG (1988): Organizing for results: The basis of authentic restructuring and reform. Educ Leadersh 46, 4–8

Spady WG: Outcome-based education: Critical issues and answers. American Association of School Administrators, Arlington 1994

Spiel C, Gössler PM (2000): Zum Einfluß von Biasvariablen auf die Bewertung universitärer Lehre durch Studierende. Z Padagog Psycol 14, 38–47

Swets JA, Dawes RM, Monahan J (2000): Better decisions through science. Sci Am 283, 82–87

Tavakol M, Dennick R (2011): Making sense of Cronbach's alpha. Int J Med Educ 2, 53–55

Theall M, Franklin J (2001): Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? New Directions for Institutional Research 109, 45–56

Thomas PA, Kern DE, Hughes MT, Chen BY (Hrsg.): Curriculum Development for Medical Education: A Six-Step Approach. 3. Auflage; Johns Hopkins University Press, Baltimore 2016

Thompson BM, Rogers JC (2008): Exploring the learning curve in medical education: Using self-assessment as a measure of learning. Acad Med 83, 86–88

VanGeest JB, Johnson TP, Welch VL (2007): Methodologies for improving response rates in surveys of physicians: A systematic review. Eval Health Prof 30, 303–321

Wachtel HK (1998): Student evaluation of college teaching effectiveness: a brief review. Assess Eval High Educ 23, 191–212

Ward M, Gruppen L, Regehr G (2002): Measuring self-assessment: Current state of the art. Adv Health Sci Educ Theory Pract 7, 63–80

Wass V, van der Vleuten C, Shatzer J, Jones R (2001): Assessment of clinical competence. Lancet 357, 945–949

Woloschuk W, Coderre S, Wright B, McLaughlin K (2011): What factors affect students' overall ratings of a course? Acad Med 86, 640–643

Zweig MH, Campbell G (1993): Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. Clin Chem 39, 561–577

Internetquellen:

BAG: Leitfaden für die Planung von Projekt- und Programmevaluation. http://www.degeval.de/fileadmin/user_upload/Sonstiges/leitfaden.pdf, abgerufen am: 27.07.2020

Benton SL, Cashin WE: IDEA Paper No. 50: Student Ratings of Teaching: A Summary of Research and Literature. https://ideacontent.blob.core.windows.net/content/sites/2/2020/01/PaperIDEA_50.pdf, abgerufen am: 27.07.2020

DeGEval: Standards für Evaluation. https://www.degeval.org/fileadmin/Publikationen/DeGEval-Standards_fuer_Evaluation.pdf, abgerufen am: 27.07.2020

Göttinger Lernzielkatalog. https://www.umg.eu/fileadmin/user_upload/Go__ttinger_Lernzielkatalog_Stand_19-07-2017.pdf, abgerufen am: 27.07.2020

Gravestock P, Gregor-Greenleaf E: Student course evaluations: Research, models and trends. The Higher Education Quality Council of Ontario. https://heqco.ca/wp-content/uploads/2020/03/Student-Course-Evaluations_Research-Models-and-Trends.pdf, abgerufen am: 11.07.2021

IMPP – Zusammenstellung der Prüfungsinhalte für den Zweiten Abschnitt der Ärztlichen Prüfung. https://www.impp.de/blueprint-m2-examen.html?file=files/PDF/Allgemein/Blueprint%20M2-Examen.pdf, abgerufen am: 27.07.2020

MFT Medizinischer Fakultätentag der Bundesrepublik Deutschland e. V.: Nationaler Kompetenzbasierter Lernzielkatalog Medizin. https://medizinische-fakultaeten.de/wp-content/uploads/2021/06/nklm_final_2015-12-04.pdf, abgerufen am: 11.07.2021

Raupach und Schiekirka – User manual for the CSA gain evaluation tool. https://www.scantron.com/wp-content/uploads/2018/09/CSA-gain-evaluation-tool_CC.pdf, abgerufen am: 27.07.2020

Senat der Georg-August-Universität Göttingen: Ordnung über die Evaluation der Lehre. https://www.uni-goettingen.de/de/evaluationsordnung+%28pdf%29/496181.html, abgerufen am: 27.07.2020

Wissenschaftsrat: Empfehlungen zur Qualitätsverbesserung von Lehre und Studium. https://www.wissenschaftsrat.de/download/archiv/8639-08.pdf?__blob=publicationFile&v=2, abgerufen am: 27.07.2020

Danksagung

Ich möchte mich an dieser Stelle ganz herzlich bei meinem Doktorvater Prof. Dr. med. Tobias Raupach für die Überlassung des Themas und die Unterstützung während der gesamten Zeit bedanken.

Des Weiteren gilt mein Dank Herrn PD Dr. med. Christian von der Brelie als Zweitbetreuer meiner Dissertation sowie Herrn Prof. Dr. Tim Beißbarth für die Unterstützung bei der statistischen Analyse der Daten.

Darüber hinaus möchte ich mich bei allen Mitarbeiterinnen und Mitarbeitern sowie Doktorandinnen und Doktoranden der Abteilung für Medizindidaktik und Ausbildungsforschung der Universitätsmedizin Göttingen für die guten Ratschläge sowie die hilfreichen Diskussionen bedanken, die einen wertvollen Teil zu dieser Arbeit beigetragen haben. Ein besonderer Dank gilt hier Janina Barth für die Unterstützung bei der Durchführung und Auswertung der Studie. Außerdem möchte ich mich herzlich bei Julia Henninger bedanken, deren Unterstützung während des Verfassens der Dissertation mir eine große Hilfe war.