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

Bedeutung von Elaboration und Kontrastierung von Falschantworten für die Entwicklung differentialdiagnostischer und -therapeutischer Entscheidungskompetenz im Medizinstudium

INAUGURAL-DISSERTATION

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

vorgelegt von

Milena Maria Berens, geb. Goldmann

aus

Hamburg

Göttingen 2022

Dekan:	Prof. Dr. med. W. Brück

Betreuungsausschuss

Betreuer/in	Prof. Dr. med. T. Raupach, MME
Ko-Betreuer/in:	Prof. Dr. med. M. Koziolek

Prüfungskommission

Referent/in	Prof. Dr. med. T. Raupach, MME
Ko-Referent/in:	..…….
Drittreferent/in:	…..…….

Datum der mündlichen Prüfung: ……..

Hiermit erkläre ich, die Dissertation mit dem Titel „Bedeutung von Elaboration und Kontrastierung von Falschantworten für die Entwicklung differentialdiagnostischer und -therapeutischer Entscheidungskompetenz im Medizinstudium“ eigenständig angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Göttingen, den …… ……

(Unterschrift)

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

Becker T, Berens M, Raupach T: Digitale formative Key-Feature-Prüfungen im Medizinstudium: Ein innovatives und evidenzbasiertes Lehrformat zur Vermittlung klinischer Entscheidungskompetenz. In: Leben N, Reinecke K, Sonntag U (Hrsg.): Hochschullehre als Gemeinschaftsaufgabe. Akteur:innen und Fachkulturen in der lernenden Organisation. wbv Publikation, Bielefeld 2022, 41-45

Goldmann M, Hasenfuß G, Dehl T, Raupach T (2016): Klug entscheiden: . . . auch in der Lehre! Dtsch Arztebl Int 113, A-2149-2154

Goldmann M, Middeke A-C, Schuelper N, Dehl T, Raupach T (2017): Klug entscheiden in der Lehre. Z Evid Fortbild Qual Gesundhwes 129, 22–26

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

CR = *Clinical Reasoning*

EE1 = *Entry Exam*

EE2 = *Exit Exam*

EFS = E-Fallseminar

KCR = *Knowledge of Correct Result*

KF = *Key Feature*

KFP = *Key Feature*-Prüfung

MCQ = *Multiple Choice Question*

PMP = *Patient Management Problem*

RT = *Retention Test*

UMG = Universitätsmedizin Göttingen

1 Einleitung

1.1 Medizinstudium in Deutschland

Die medizinische Ausbildung in Deutschland besteht aus einem sechsjährigen Universitätsstudium, welches 5500 Stunden Unterrichtszeit umfasst (ÄApprO 2002) sowie einen dreimonatigen Krankenpflegedienst, eine Famulatur von vier Monaten und eine Ausbildung in Erster Hilfe. Das letzte Jahr des Studiums sieht eine 48 Wochen dauernde praktische Ausbildung vor (Praktisches Jahr). Zum Erwerb des Staatsexamens muss die Ärztliche Prüfung abgelegt werden. Diese ist in drei Abschnitte aufgeteilt – zwei schriftliche und einen mündlich-praktischen – und ermöglicht nach ihrem Bestehen die Beantragung der Approbation als Arzt. Mit Erhalt der Approbation darf ein Arzt[1] praktizieren und sich zum Facharzt weiterbilden lassen. Die Anerkennung einer Facharztweiterbildung setzt eine mehrjährige Tätigkeit und Ausbildung im entsprechenden Fachbereich und das Bestehen einer Prüfung voraus. Auch nach abgeschlossener Weiterbildung besteht laut der Musterberufsordnung für Ärzte für jeden praktizierenden Mediziner die Verpflichtung, sich beruflich fortzubilden (Bundesärztekammer 2019). Somit legt das Medizinstudium den Grundstein für eine ärztliche Tätigkeit, die von einem kontinuierlichen Lernprozess begleitet wird. Ziel der universitären Ausbildung ist es, die grundsätzlich notwendigen Kenntnisse und Fähigkeiten zur ärztlichen Berufsausübung zu vermitteln, aber auch die Befähigung zur Weiter- und Fortbildung (ÄApprO 2002). Dieser bereits in Paragraph 1 der Approbationsordnung für Ärzte erhobene Anspruch belegt die große Bedeutung des lebenslangen Lernens für den Arztberuf.

Bereits zu Beginn des Medizinstudiums muss meist eine große Menge an Lernstoff bewältigt werden. Für jeden Erkenntnisschritt müssen zunächst die Grundlagen erarbeitet werden, auf denen alle späteren Überlegungen fußen. Um Konzepte der Krankheitsentstehung und Heilung verstehen zu können, müssen Aufbau und Physiologie des menschlichen Körpers verstanden werden. Um auch komplexe Vorgänge auf zellulärer und molekularer Ebene einbeziehen zu können, müssen die zugrundeliegenden naturwissenschaftlichen Prinzipien beherrscht werden. Studierenden der Medizin wird dadurch ein ausgedehntes Lernpensum auferlegt, um die zahlreichen Ebenen des Verständnisses aufzubauen, welche später bei der Lösung eines medizinischen Problems helfen sollen. Durch immer neue wis-

[1] In dieser Arbeit werden bei allen Bezeichnungen von Berufen, Gruppen, Positionen und sozialen Rollen stets beide Geschlechter inbegriffen. Aus Gründen der Lesbarkeit wird das generische Maskulinum verwendet und auf die Verwendung von Sonderzeichen zur geschlechtergerechten Kennzeichnung verzichtet, gemeint sind dabei aber ausdrücklich und uneingeschränkt gleichberechtigt beide Geschlechter.

senschaftliche Erkenntnisse und die daraus resultierenden zahlreicheren medizinischen Möglichkeiten erscheint das zu lernende Wissen vielfältiger denn je. Um diesen hohen Anforderungen gerecht werden zu können, müssen Medizinstudierende effektive Lernmethoden anwenden, die das sowohl im Selbststudium als auch im Präsenzunterricht angeeignete Wissen festigen. Im Verlauf des Studiums wird außerdem zunehmend die Übertragung von Faktenwissen auf individuelle Situationen wichtig. Nach der Aneignung eines breiten Grundwissens muss der Studierende also auch dessen Verknüpfung und Anwendung lernen. Da das Ziel die Befähigung zur ärztlichen Tätigkeit ist, muss der Studierende lernen, wie ein Arzt zu denken.

1.2 *Clinical Reasoning* – wie Ärzte denken

Ärzte treffen in ihrem beruflichen Alltag auf eine Vielzahl von unterschiedlichen Patienten und Situationen, meist innerhalb von kurzer Zeit. Zunächst muss ein Fall in seinem Kontext erfasst und eine klinische Hypothese festgelegt werden. Allein schon darin liegt durch die Individualität jedes Patienten und aufgrund der komplexen Wechselwirkungen zwischen dem Menschen, seiner Umwelt, seinem Selbstbild und nicht zuletzt auch mit seinem Arzt eine Herausforderung (Beach et al. 2006; Faller und Lang 2011). In manchen Fällen müssen Ursachen für ein Problem gefunden werden, in anderen Fällen mögliche Lösungen. Der Arzt kann sich in einer Routinesituation befinden, aber auch mit Notfällen und Ausnahmesituationen konfrontiert werden.

Im Klinikalltag werden daher von einem Arzt unterschiedliche Kompetenzen verlangt: Soziale Fähigkeiten sind für den Umgang mit Patienten und Arbeitskollegen wertvoll, Zeitmanagement und organisatorisches Geschick bei der Verteilung und Priorisierung von Aufgaben sind im schnelllebigen und unvorhersehbaren Bereich der Patientenversorgung notwendig (Fournier 2000), zuweilen auch handwerkliche Geschicklichkeit, technisches Know-How und feine Sinnesorgane (Schnabel et al. 2011). Das Spektrum der von einem Arzt geforderten Kompetenzen ist so vielfältig wie die zahlreichen medizinischen Fachdisziplinen (MFT Medizinischer Fakultätentag der Bundesrepublik Deutschland e. V. 2015). Zu jeder Zeit und von jedem Mediziner wird jedoch eine Übertragungsleistung von theoretischem Wissen und Erfahrungen in die konkrete Praxis gefordert, ist das Anwenden von abgespeicherter Information tragende Einflussgröße beim Treffen von Entscheidungen.

Während also die Bedingungen, unter denen Ärzte auf ihre Patienten treffen, sehr verschieden sein können, treffen einige Anforderungen in jedem Fall ärztlichen Handelns zu. Ob im persönlichen Kontakt, im Austausch mit anderen Experten oder beim Verfassen von Arztbriefen, es ist immer notwendig, sich auf den individuellen Fall einzustellen, der bei keinen zwei Patienten der gleiche ist. Es muss also stets eine Moderation zwischen dem Wissen über die Allgemeinheit und über den Einzelfall stattfinden. Während lange Zeit über im Medizinstudium Grundwissen aus Listen und Büchern gelernt wird, findet sich der Arzt bei seiner Arbeit vor ganz anderen kognitiven Herausforderungen wieder. Nachdem

der angehende Mediziner sich zunächst angeeignet hat, wie der menschliche Körper funktioniert oder eben nicht funktioniert, muss er lernen, wie ein Arzt Probleme löst.

Die kognitiven Prozesse, welche der ärztlichen Entscheidungsfindung im Rahmen diagnostischer und therapeutischer Überlegungen zugrunde liegen, werden klinisches Denken, auf Englisch *Clinical Reasoning* (CR), oder auch *Clinical Decision Making* genannt (Eva 2005), in der Literatur tauchen aber auch Begriffe wie *Diagnostic reasoning* (Bowen 2006), *Clinical cognition* (Kassirer 2010), *Medical decision making* (Patel et al. 2002) oder *Clinical judgment* (Croskerry 2009) auf. Viele haben versucht, diese Begriffe genauer zu definieren, die meist synonym gebraucht werden, es existieren daher zahlreiche Erklärungsansätze.

1.2.1 Definition und Bedeutung

Clinical Reasoning wird beschrieben als die Fähigkeit, die geballten von einem Patienten präsentierten Merkmale zu ordnen und mit einem korrekten Diagnose-Label zu versehen, mit dem Ziel der Entwicklung einer angemessenen Behandlungsstrategie (Eva 2005). Klinisches Denken umfasst dabei die Auswahl an Strategien, die Kliniker benutzen, um Diagnosen zu generieren, zu testen und zu verifizieren, um Nutzen und Risiken von Untersuchungen und Therapien zu bewerten und um die prognostische Bedeutung der Ergebnisse dieser Überlegungen zu beurteilen (Kassirer 2010). Es ist außerdem die Fähigkeit, verschiedene Arten von Wissen anzuwenden, Befunde kritisch zu betrachten und den eigenen Vorgang bei der Diagnosefindung zu reflektieren (Modi et al. 2015). Ähnlich wie die beiden hier zitierten ranken sich weitere Definitionen von *Clinical Reasoning* um den Prozess der kognitiven Bearbeitung eines Patientenfalles durch den Arzt mit allen Überlegungsschritten auf dem Weg zur richtigen Diagnose und Behandlung (siehe z.B. Gruppen 2017). Der Begriff umfasst also – vereinfacht ausgedrückt – das, was im Kopf eines Behandlers vorgeht, wenn er Patientenfälle erfasst und klinische Entscheidungen trifft. In diesem breiten Sinne soll *Clinical Reasoning*, beziehungsweise klinisches Denken, in dieser Arbeit verstanden werden. Diese vage und schwer einzugrenzende Definition von *Clinical Reasoning* entspricht dabei genau der Vielgestaltigkeit dieses Begriffs, die im Folgenden näher beleuchtet wird.

So unscharf diese Definition des klinischen Denkens zunächst wirken mag, herrscht bei allen Definitionen Einigkeit darüber, dass es eine sehr wichtige Rolle in der ärztlichen Tätigkeit spielt. Die klinische Urteilsfähigkeit ist als entscheidender Aspekt ärztlicher Arbeitsleistung essentiell für die Formulierung einer Diagnose und der Schlüssel für effektives und sicheres Patientenmanagement (Croskerry 2009). *Clinical Reasoning* wird in der Literatur oft als Kernkompetenz bezeichnet, deren Aneignung von allen Klinikern, unabhängig von ihrem Fachgebiet, erwartet wird (Norman 2005; Cutrer et al. 2013; Modi et al. 2015) – die zugeschriebene Bedeutung reicht bis hin zum definierenden Charakteristikum der medizinischen Profession (Gruppen 2017). Wenn diese Fähigkeit tatsächlich so entscheidend ist, leuchtet es ein sich zu fragen, was sie ausmacht. Wie genau funktioniert klinisches Denken und was kann man tun, um es zu lernen?

1.2.2 Funktionsweise und Modellierung

Unter dem Begriff *Clinical Reasoning* wird meist ein komplexes Zusammenspiel aus Wissen, Erfahrung und deren Integration mit vorliegenden Patienteninformationen zu einer abstrahiert-strukturierten Fallrepräsentation subsummiert, welche als gedankliche Leitschiene bei der Planung des weiteren klinischen Vorgehens dient. Während des Zugewinns weiterer Informationen wird dieses gedankliche Konzept stetig überprüft und erneuert, was zu weiteren Schleifen von Informationsgewinnung und Überprüfung des Fallkonzeptes führt (Kassirer 2010; Gruppen 2017). Allen Erklärungsansätzen gemein ist, dass *Clinical Reasoning* als ein komplexer, vielschichtiger kognitiver Prozess eines Klinikers bei der Auseinandersetzung mit einer medizinischen Problemstellung verstanden wird, der mehrschrittig, andauernd und auf verschiedenen Ebenen gleichzeitig stattfindet und der zwar in Ansätzen gut charakterisiert, in seiner Gänze jedoch aus einer einzigen Perspektive kaum erfasst werden kann.

Nach Norman (2005) ist *Clinical Reasoning* kein einheitlicher Prozess, der in seine Einzelschritte zerlegt und wie andere Fertigkeiten nach einem bestimmten Schema gelernt werden kann. Es besteht vielmehr aus einer Vielzahl unterschiedlicher kognitiver Mechanismen, denen sich der Arzt interindividuell unterschiedlich und situationsangepasst bedient.

Frühere Arbeiten mit der Fragestellung, durch welche gedanklichen Prozesse sich klinisches Denken auszeichnet, konnten zunächst keinen eindeutigen Mechanismus identifizieren, den erfahrene Kliniker im Gegensatz zu Anfängern anwendeten, um klinische Probleme zu lösen (Elstein et al. 1978; Neufeld et al. 1981). In der groben Struktur des Lösungsprozesses waren keine großen Unterschiede erkenntlich. Stattdessen schien die Findung des richtigen Ergebnisses, zum Beispiel bei der Diagnosestellung, vom Wissensstand des getesteten Klinikers auf dem jeweiligen Gebiet abzuhängen. Dass klinisches Denken nur eine Frage des Gedächtnisses und der Menge abrufbaren spezifischen Wissens sei, konnte in darauf folgenden Untersuchungen jedoch nicht gezeigt werden (Muzzin et al. 1983; Norman et al. 1985; Patel et al. 1986). Im Gegenteil wurde sogar unter dem Begriff „Intermediate Effect“ beschrieben, dass fortgeschrittene Studenten und junge Ärzte in einigen Domänen, die zur Abbildung des klinischen Denkens geprüft wurden, mehr Leistung zeigten als erfahrene Ärzte: beispielsweise wurden von ihnen mehr Informationen aus Patientenfällen behalten und ausführlichere pathophysiologische Erklärungen für Symptome abgegeben (Schmidt und Boshuizen 1993; Rikers et al. 2000). Mediziner gegen Ende des Studiums und Berufsanfänger griffen laut Grant und Marsden auf eine einheitlichere Basis an Grundwissen zurück, welche bei erfahreneren Ärzten zugunsten individueller, auf Erfahrung basierender Kenntnisse in den Hintergrund trat. Die Menge an Wissen, auf das beim Lösen klinischer Fälle tatsächlich zurückgegriffen wurde, stieg nicht linear mit zunehmender klinischer Erfahrung an (Grant und Marsden 1988). Sowohl der alleinige Denkprozess als auch alleiniges Fachwissen lieferten also keine hinreichende Erklärung zur Charakterisierung des Wesens erfolgreicher klinischer Entscheidungsfindung. Schmidt und Boshuizen

stellten aufgrund ihrer Beobachtungen, dass erfahrene Kliniker in den meisten Fällen zur Problemlösung nicht auf Grundlagenwissen zurückgreifen, die Theorie auf, dass das zugrundeliegende naturwissenschaftliche Basiswissen „eingekapselt" im Gedächtnis des Experten vorliegt, sodass es nicht vordergründig am Denkprozess beteiligt ist, er aber, wenn nötig, darauf zurückgreifen kann (Schmidt und Boshuizen 1993). Das Grundlagenwissen wird nach diesem Modell implizit in klinisches Wissen integriert. Diese Theorie konnte empirisch belegt werden (Bruin et al. 2005) und kommt bis heute im wissenschaftlichen Diskurs vor (Schmidt und Mamede 2015). Obwohl der Experte also über mehr Basiswissen in seinem Themengebiet verfügt, wird dieses seltener gebraucht und vornehmlich zur Lösung schwierigerer oder komplexerer Probleme mobilisiert (Norman 2005).

Nicht die Menge an Interpretationen und Hypothesen unterscheide sich zwischen Anfängern und Fortgeschrittenen, sondern deren Genauigkeit:

> *„We conclude that there is no difference between groups of differing clinical experience in the breadth of thought but that there are marked differences in the precise content and structure of thought."* (Grant und Marsden 1987)

Wenn es also nicht die grobe Herangehensweise ist, so müssen es zumindest unterbewusste Mechanismen im Denkprozess sein, die gekonntes klinisches Denken ausmachen. Da es unstrittig erscheint, dass Wissen eine wichtige Rolle für das klinische Denken spielt, ohne dass Gedächtnis und Erinnerungsfähigkeit alleinige Determinanten für gute Leistung wären, wurde untersucht, wie die unterschiedlichen Arten von Wissen im Gedächtnis des Mediziners organisiert und strukturiert werden – also wie und in welcher Form Wissen greifbar gemacht und angewandt wird.

Dazu gibt es verschiedene Konzepte:

Eine Theorie für die Entwicklung kompetenten klinischen Denkens benennt kognitive Strukturen, die prototypische Patientenfälle darstellen, sogenannte *Illness Scripts* (in etwa: krankheitsspezifische Manuskripte). Diese innere Repräsentation einer Krankheit als prototypischer Patient, die aus der Erfahrung des Klinikers entstanden ist, ist direkt verknüpft mit klinisch relevanten Informationen, weniger mit pathophysiologischen Herleitungen. Laut Schmidt et al. durchlaufen Mediziner im Zuge von Studium und Weiterbildung mehrere Stadien der unterschiedlichen Wissensstrukturierung bei der Anwendung von Wissen in klinischen Denkprozessen. Die Nutzung von Prototypen sei assoziiert mit fortgeschrittener klinischer Erfahrung, Studenten ohne klinische Erfahrung benutzten kausale pathophysiologische Krankheitsmodelle (Schmidt et al. 1990).

Aufbauend auf dieser Annahme, Diagnosefindung sei für den Arzt in Routinesituationen ein Mechanismus der Kategorisierung und Mustererkennung, und daraus folgend, dass die Leistung auf dem Vergleich von Fallbeispiel und mentalen Prototypen basiere, wurde die Fähigkeit, diese Muster voneinander zu unterscheiden, als Prädiktor für die Treffsicherheit von Diagnosen experimentell belegt (Papa et al. 1990). Nicht nur das Vorhandensein von

Mustern, sondern vor allem deren Diskriminierung spielt also eine wichtige Rolle im Prozess der Entscheidungsfindung. Je deutlicher unterschiedlich die Prototypen voneinander seien, desto eher werden Fallbeispiele korrekt klassifiziert (Papa et al. 1990).

Für das von Papa et al. durchgeführte Experiment wurde ein auf künstlicher Intelligenz basierendes Programm benutzt, das aus geschätzten Wahrscheinlichkeiten, wie häufig bestimmte Befunde bei bestimmten Krankheiten auftreten, krankheitsspezifische Muster erstellt. Der Prozess des klinischen Denkens wurde hier also als ein Abwägen kombinierter Wahrscheinlichkeiten modelliert. Dass dies funktionieren kann, zeigte das Modell, dennoch stellt sich die Frage, wie klinisches Denken in der Praxis tatsächlich abläuft.

Bordage und Lemieux beschreiben in ihren Arbeiten semantische Netzwerke, bestehend aus qualitativ beschreibenden Begriffen, deren Benutzung im diagnostischen Denkprozess mit mehr Erfolg bei der Diagnosestellung einhergeht (Bordage und Lemieux 1991). Diese Begriffe, die bei der Interpretation und Kommunikation von Patientenfällen zum Tragen kommen, treten meist als oppositionelle Paare auf – beispielsweise „akut“ und „chronisch“ oder „anhaltend“ und „intermittierend“ – und bilden jeweils eine bipolare Achse einer bestimmten Eigenschaft. Sowohl Medizinstudenten als auch Fachärzte, die besonders gute Ergebnisse bei der Bearbeitung von Patientenfällen zeigten, zeichneten sich durch die Benutzung besonders vieler dieser Begriffe aus (Bordage und Lemieux 1991; Chang et al. 1998). Die Abstrahierung von Informationen und ihre Einordnung in ein multiaxiales System diagnostischer Qualitäten wird mit dem Vorhandensein einer tiefergehenden mentalen Fallrepräsentation in Verbindung gebracht, in der ein Patientenfall schematisch-strukturiert beschrieben und analysiert wird, im Gegensatz zu einer unstrukturierten Generierung von Verdachtsdiagnosen. Durch die klare Einordnung auf möglichst vielen Achsen entsteht ein Bild, anhand dessen Differentialdiagnosen aus- oder eingeschlossen werden können.

Dass die Fähigkeit, etwas präzise zu beschreiben, mit tieferem Verständnis des Inhalts zusammenhängt, erscheint nachvollziehbar. Die Benutzung von definierten abstrahierenden Begriffen setzt ein klares inneres Bild von deren Bedeutung voraus. Für die Zuordnung von diagnostischen Labels, wie bei der Anwendung der oben beschriebenen bipolaren Qualitäten, wird klinisches Wissen (zum Beispiel über Bedeutung und typisches Erscheinungsbild von klinischen Eigenschaften) benötigt, andererseits helfen Kenntnis und Verwendung dieser Begriffe bei der Aneignung klinischen Wissens, beispielsweise wenn Fälle auf diese Weise strukturiert und kategorisiert abgespeichert werden können.

Bisher wurden verschiedene Strategien vorgestellt, für die belegt wurde, dass sie im Rahmen von *Clinical Reasoning* stattfinden. Aber wann kommt welche Denkweise zum Tragen? Das als nächstes beschriebene Konzept der Einteilung dieser Strategien dient der vereinfachten Vorstellbarkeit der Abläufe beim klinischen Denken.

Der *Clinical Reasoning*-Prozess wird häufig als ein duales System beschrieben, bestehend aus intuitiven Komponenten einerseits und analytischen Komponenten andererseits (Kassirer 2010). Dieses duale System des Denkens wurde zuvor beschrieben von Daniel Kahneman

(Kahneman 2011). Die intuitiven Komponenten laufen schnell und oft unterbewusst ab; dazu gehören Mustererkennung, erste Eindrücke, automatische Einordnung in Kategorien und Analogien zu früheren Fällen. Diese reflexartig und instinktbasiert ablaufenden Automatismen dienen der schnellen Einordnung und Bewertung neuer Situationen, sind eher unwillkürliche Reaktionen und benötigen keine bewussten Denkanstrengungen. Die intuitiven Prozesse sind geprägt durch Erfahrung und beeinflussbar durch Emotionen und situative Stimmung. Demgegenüber stehen die analytischen Komponenten: bewusste Denkprozesse, bei denen gezielt Fakten und Optionen abgewogen werden. In diese Kategorie fallen differentialdiagnostische Erwägungen, pathophysiologische Erklärungsansätze, gezielte Hypothesenverifikation und –testung, die Anwendung von diagnostischen und therapeutischen Algorithmen. Beim analytischen Denken werden Befunde kritisch betrachtet, mit abgespeichertem Wissen in Zusammenhang gebracht, Schritt für Schritt Erklärungen formuliert und bisherige Schritte reflektiert. Es ist langsamer als die intuitiven Komponenten und erfordert mehr kognitive Anstrengung. Im Gegensatz zu Intuition und Gefühl basiert die analytische Komponente auf Logik, Wissen und bewusster Entscheidung. Analytische Schlussfolgerungen können auf Kausalität (zum Beispiel pathophysiologische Erklärbarkeit) oder auf Wahrscheinlichkeitsmodellen beruhen (Eva 2005; Kassirer 2010).

Die beiden Komponenten klinischen Denkens komplementieren einander. Mit schnellen ersten Reaktionen und Eindrücken kann sofort zu arbeiten begonnen werden, bis sich im Idealfall Zeitpunkte ergeben, die daraus gezogenen Schlüsse zu überdenken und analytisch nachzuvollziehen. Gerade die intuitiven Anteile klinischen Denkens sind – beispielsweise aufgrund ihrer Beeinflussbarkeit durch Kontextfaktoren – fehleranfällig. Jedoch sollte auch das Fehlerpotential in langsamen analytischen Prozessen nicht unterschätzt werden, genauso wie der Wert einer starken Intuition (Kassirer 2010). Fälle schnell richtig einschätzen und instinktiv handeln zu können, ist eine bemerkenswerte Fähigkeit erfahrener Kliniker, die im Alltag eine große Hilfe darstellt. Ohne lange nachdenken zu müssen die richtige Entscheidung zu treffen, ist effizient und kann zurecht als ein Ziel guter Aus- und Weiterbildung gesehen werden. Genauso sollte ein Experte jedoch sorgfältig analytisch vorgehen können, um den Fällen gerecht zu werden, die nicht ins übliche Bild passen. In seinem Review *Thinking about diagnostic thinking: a 30-year perspective* sieht Arthur Elstein das Problem in der Frage, wann der Arzt auf welche Art des Denkens zurückgreift und wie er entscheidet, wann beispielsweise ein langsamerer, sorgfältigerer Prozess notwendig wird (Elstein 2009). Es wird angenommen, dass, wer unerfahren ist, eher auf analytische Strategien zurückgreift, bis er über ausführlicheres Erfahrungswissen verfügt. Durch Übung kann ein vormals analytischer Prozess automatisiert werden. Routinefälle werden so eher durch schnelle Mustererkennungsmechanismen gelöst, während für schwierigere Fälle weiterhin analytische Strategien benötigt werden (Kassirer 2010). Letztlich kommen beide Arten von *Clinical Reasoning* bei erfahrenen und unerfahrenen Klinikern zur Anwendung (Eva 2005; Norman 2005; Elstein 2009). Obwohl die analytischen und die intuitiven Komponenten so gegensätzlich erscheinen, sind sie in der Realität vermutlich nicht vollends voneinander

trennbar, wenn nicht sogar miteinander verflochten. Nicht jedes medizinische Denken lässt sich eindeutig einer der beiden Kategorien (Croskerry 2009) zuordnen. Kevin Eva erklärt das Verhältnis der beiden Denkweisen folgendermaßen:

> *„these two forms of reasoning should be viewed as being very interactive; rather than lying along a continuum, they are instead complementary contributors to the overall accuracy of the Clinical Reasoning process, each influencing the other"* (Eva 2005).

Wie nun deutlich geworden ist, besteht klinisches Denken aus unterschiedlichen Komponenten, die im besten Fall situationsangepasst flexibel zum Einsatz kommen. Obwohl bereits viel über *Clinical Reasoning* geforscht wurde, kann das Prinzip nur in Teilen erfasst werden. Es existieren ausgearbeitete, teils aufwendige Modelle über die genauen kognitiven Vorgänge (Eva 2005; Croskerry 2009; Charlin et al. 2012), diese bleiben als Modelle jedoch Annäherungsversuche – mit dem erklärten Ziel, eine Hilfe zur besseren Vorstellung zu sein, um ausreichend Verständnis von *Clinical Reasoning* nicht nur aus wissenschaftlichem Interesse, sondern zu Lern-, Lehr- und Prüfungszwecken zu ermöglichen. Denn so komplex und schwer zu fassen *Clinical Reasoning* auch ist, Einigkeit herrscht darüber, dass angehende Ärzte es lernen und üben müssen und dass dieses Lernziel geprüft werden muss (Norman 2005).

1.3 Lernen von *Clinical Reasoning*

Wie kann man lernen, klinisch zu denken? Die größte Motivation für die Erforschung des Wesens von *Clinical Reasoning* ist das Ziel, gut ausgebildete Ärzte in den Praxen und Krankenhäusern zu haben, die richtige Diagnosen stellen und für ihre Patienten die richtigen Entscheidungen treffen. Grundlage für die Konzeption einer effektiven Lernstrategie ist das Verständnis der zugrundeliegenden Prozesse und Abläufe beim Lösen eines klinischen Falls. Mit einem guten Modell des zu lernenden können sowohl Lehrer als auch Lernende ausgestattet werden, um besser zum Ziel zu kommen. Auch für Metakognition zur Selbstüberprüfung des eigenen Lösungsprozesses ist eine Vorstellung von dessen Funktionsweise wichtig. Es wird darauf hingewiesen, dass ein unvollständiges Verständnis von *Clinical Reasoning* ein bedeutendes Hindernis für die Effektivität in der medizinischen Lehre darstellt. (Groves 2012). Ebenso braucht es wirksame Methoden, die das Erlernen von klinischer Problemlösung ermöglichen. Diese sollten passend auf dieses Lernziel ausgelegt sein. Zwar kann *Clinical Reasoning* auch beim Lernen von Fakten und sammeln von Erfahrung „nebenbei" erworben werden, jedoch fordert deren umfassender Stellenwert in der ärztlichen Kunst eine explizitere Aufmerksamkeit in der Lehre (Groves 2011; Schmidt und Mamede 2015).

1.3.1 Elaboration

Bevor es um Methoden zum Lernen von *Clinical Reasoning* geht, soll an dieser Stelle der lerntheoretische Begriff „Elaboration" eingeführt werden, welcher eine Schnittstelle zwischen kognitiven Prozessen des *Clinical Reasoning* und des Lernens bildet.

Wie bereits im vorherigen Kapitel im Zusammenhang mit den semantischen Netzwerken von Bordage und Lemieux beschrieben, scheint das in-Worte-fassen eines Sachverhaltes mit Lernen und Verständnis dieses Inhaltes zusammenzuhängen. Dass es beim Lernen hilft, etwas erklären zu können – und insbesondere, dies auch tatsächlich zu tun – ist als generative Lernstrategie anerkannt (Hasselhorn und Gold 2009; Weinstein et al. 2011) und kann durch den zugrundeliegenden Prozess der Elaboration erklärt werden. Elaboration beschreibt eine kognitive Umstrukturierung des Lernmaterials, die für die Verknüpfung und Abspeicherung von Informationen im Gedächtnis benötigt wird (Slavin 1996). Dazu gehört z.B. die Schaffung von Bedeutung durch Konstruktionen und Ergänzungen (Levin 1988) und die Verknüpfung mit bereits vorhandenem Wissen (Weinstein et al. 2011). Das Lernmaterial jemand anderem zu erklären wird als eine der effektivsten Arten von Elaboration bezeichnet (Slavin 1996). Somit gibt es gute Hinweise darauf, dass durch Elaboration, beispielsweise beim konkreten Erläutern und Herleiten von Lerninhalten, eine bessere Retention der Lerninhalte im Gedächtnis erreicht werden kann. Dies ist wichtig für das Lernen von *Clinical Reasoning* angesichts der Fülle an Informationen, die verknüpft und erinnert werden müssen.

1.3.2 Lernen anhand von Fallbeispielen

Weiterhin weckt die Frage, durch welche Lehrformen und Curricula klinisches Denken am besten vermittelt werden kann, viele Bestrebungen in der medizinischen Lehrforschung und auch hier sind die Erkenntnisse divers – mit einem gemeinsamen Nenner, auf dem auch das in dieser Arbeit untersuchte Lernformat aufgebaut ist.

Dieser gemeinsame Nenner ist Lernen anhand von Fallbeispielen. Einerseits sind Kurse, in denen klinisches Denken allgemein als eine Fertigkeit oder als ein Schritt-für-Schritt-Prozess beigebracht wird, aufgrund der fehlenden Evidenz für die Wirksamkeit solcher Ansätze und für die Existenz einer solchen allgemeinen Fähigkeit, nicht sinnvoll (Schmidt und Mamede 2015). Es wird nicht mehr davon ausgegangen, dass *Clinical Reasoning* als solches wie ein Kochrezept für alle klinischen Situationen, Anwender und Fächer beigebracht werden kann. Andererseits wird durch das Lernen an Fallbeispielen das zu Lernende direkt in dem Kontext erfahren, in dem es später angewendet werden soll. Dies bietet den Vorteil, dass *Clinical Reasoning* so direkt trainiert werden kann und dieses Training soll letztlich den Lerneffekt, die Aneignung dieser komplexen Problemlösefähigkeit, bewirken. Obwohl es auf jedem Erfahrungslevel gute und weniger gute klinische Denker gibt, unterscheiden sich Kliniker mit exzellenter *Clinical Reasoning*-Kompetenz in der Regel durch umfassendere Praxiserfahrung von weniger erfolgreichen Diagnostikern (Norman 2005; Bowen 2006).

Die tragende Rolle von Übung und Erfahrung für die Entwicklung von Expertise ist der Hauptgrund, warum lernen am Fallbeispiel in der medizinischen Lehrforschung immer wieder hervorgehoben wird (Schmidt et al. 1990; Eva 2005; Norman 2005; Kassirer 2010). Um einen Erfahrungsschatz aufzubauen, aus dem sie beim Treffen späterer Entscheidungen zehren können – zum Beispiel durch das Bilden von Analogien –, müssen Studenten Patientenfällen ausgesetzt werden. Die Ansammlung von Fällen eines bestimmten Themas im Gedächtnis ermögliche den Aufbau einer Basis, von der nicht-analytische – also automatische, intuitive – Prozesse ausgehen könnten (Eva 2005). Durch ausreichend Kontakt mit Fällen einer bestimmten Krankheit können *Illness Scripts* entwickelt werden (Schmidt et al. 1990), die klinischen Eigenschaften einer Entität bekommen kognitive Struktur. Die im Gedächtnis gespeicherten *Illness Scripts* gewinnen an Ausführlichkeit und Trennschärfe, je mehr zugehörige Fälle der Lernende sieht. Man könnte auch sagen, die Vertrautheit mit der Art und Weise, wie sich eine Krankheit präsentiert und wie damit umzugehen ist, wächst. Patientenfälle können also durch die Lieferung von Beispielen beim Aufbau von Erfahrung helfen, welche benötigt wird, um solide intuitive Komponenten von *Clinical Reasoning* auszubilden.

Eng verwandt damit ist der Aspekt, dass Fallbeispiele einen Platz für Übung bieten. Im Verlauf eines klinischen Falles kommt eine Krankheit ganz anders zur Geltung als in einem ihr gewidmeten Lehrbuchkapitel. Das Nachvollziehen eines echten Verlaufes im Fallbeispiel bietet die Sichtweise eines Klinikers: Die Rolle des klinischen Denkers und Entscheiders kann eingenommen und diese zunächst ungewohnte Situation erprobt werden. Denkfehler dürfen gemacht werden, ohne dass diese direkte Konsequenzen für Patienten haben. Vielmehr können durch fehlerhafte Schlüsse oder die Unfähigkeit, überhaupt einen Schluss zu ziehen, Wissenslücken aufgedeckt und direkt bearbeitet werden.

Ein weiterer Grund, warum der Einsatz von Fallbeispielen sinnvoll ist, ergibt sich aus den oben beschriebenen Annahmen, wie *Clinical Reasoning* funktioniert: Bezogen auf die unterschiedlichen Formen von *Clinical Reasoning* heißt das, dass in der Lehre sowohl analytische als auch nicht-analytische Lösungsansätze kombiniert gelehrt werden sollen, da diese sich ergänzen und in der späteren Praxis beide benötigt werden (Eva 2005). Eine Schwäche in einer der beiden Domänen und damit eine Überbetonung der anderen führe zu schlechteren diagnostischen Ergebnissen als der kombinierte Einsatz von analytischen und intuitiven Strategien. Bei der Konfrontation mit Patientenfällen ist es möglich, sowohl schnelle, intuitive Reaktionen in Gang zu setzen und davon ausgehend erste Verdachtsdiagnosen zu generieren, als auch das langsame Prüfen und Testen zu üben.

Lehre anhand von klinischen Fällen scheint sich an medizinischen Fakultäten mehr und mehr zu konsolidieren – was sich auch in der Lehrforschung abbildet (Schmidt und Mamede 2015). Dabei geht der Trend dahin, klinisch orientierte Lehre bereits früh im Medizinstudium einzusetzen, nicht erst an dessen Ende, was auch vonseiten der Lehrforschung vorgeschlagen wird, allerdings unter der Maßgabe, dass das *Clinical Reasoning*-Training sorg-

sam auf den Stand der Studierenden zugeschnitten sein muss (Bruin et al. 2005; Kassirer 2010). Im Folgenden wird dargestellt, wie fallbasierte Lehre am besten umgesetzt werden sollte:

Da *Clinical Reasoning*-Strategien so unterschiedlich sind – von Fall zu Fall, von Situation zu Situation und von Fachgebiet zu Fachgebiet – und darüber hinaus auch die Anwender, also Studenten und Kliniker, mit ganz unterschiedlichen Erfahrungen und Denkweisen an diese Situationen herangehen, sollte sichergestellt werden, dass Studenten mit möglichst vielen und diversen Strategien ausgestattet werden, um sich flexibel an die Anfordernisse der Situation anpassen zu können (Eva 2005). Bildhaft ausgedrückt: je mehr Werkzeuge man zur Auswahl hat, desto wahrscheinlicher ist es, dass eines der Werkzeuge die Erfüllung der aktuellen Aufgabe ermöglicht (Eva 2005). Da nicht ein Lösungsweg für alle Probleme und Anwender gleichermaßen funktioniert, muss auf inhaltliche und methodische Breite gesetzt werden. Für das Unterrichten am Fallbeispiel gilt, dass Studenten möglichst viele Beispiele bekommen sollten, die in ihrer Variation die Wirklichkeit abbildeten (Eva 2005). Dazu gehört etwa, dass eine bestimmte Krankheit nicht nur anhand eines einzigen Beispiels gelernt wird, sondern anhand von möglichst vielen, die jeweils unterschiedliche Ausgangssituationen und Erscheinungsformen darstellen. Schließlich präsentieren sich nicht alle realen Patientenfälle mit allen Kriterien, die im Lehrbuch stehen. Durch die Variation können die Studenten einerseits eine mentale Datenbank von Fällen aufbauen und werden andererseits dazu angehalten, auf unterschiedlichen Wegen zum Ziel zu kommen, wodurch sie unterschiedliche Strategien trainieren können. Aufgrund der Kontextspezifität klinischer Lerninhalte sind wenige ausführliche Beispiele daher wahrscheinlich weniger optimal als viele Beispiele, die in ihrer Summe die Variationsbreite der Erscheinungsformen einer bestimmten Krankheit repräsentierten. Schließlich sind alle klinischen Fälle in ihrer Eigenart spezifisch – was auch als Fallspezifität bezeichnet wird. Das daraus gelernte soll jedoch über diese Fallspezifität hinausgehen (Eva 2005).

Außerdem soll die Art und Weise, wie die Studenten auf den Fall treffen, die Realität imitieren: die Fallgeschichte soll neu und die Diagnose nicht von vorneherein bekannt sein, sondern der Lernende soll anfangs möglichst wenig wissen. Nur so kann erprobt werden, ob er den Fall richtig einordnen kann. Darüber hinaus wird ein retrospektiver *Bias* (kognitive Verzerrung, nach Eintritt eines Ereignisses die Vorhersehbarkeit des Ereignisses zu überschätzen) verhindert (Eva 2005; Kassirer 2010). Beim Üben mit mehreren Fällen hintereinander sollte bestenfalls ein sogenannter „mixed practice"-Ansatz verfolgt werden, bei dem Fälle aus verschiedenen Kategorien vermischt vorkommen, sodass nicht anhand des Sitzungsthemas auf die zu findende Diagnose geschlossen werden kann (Eva 2005).

Auch Negativbeispiele sollten eingebunden werden – also Fallbeispiele, in denen das *Clinical Reasoning* versagt hat (Kassirer 2010). Fallerzählungen sollten neben den Daten und Fakten der Krankengeschichte auch getroffene Bewertungen und Entscheidungen enthalten, sowie die Maßnahmen, die im weiteren Verlauf ergriffen wurden (Kassirer 2010). Das

Fallmaterial soll daher in chronologischer Reihenfolge der Geschehnisse präsentiert werden. Die Fälle sollen laut Kassirer – angepasst an den Leistungsstand der Lerner – echte Patientenfälle sein (Kassirer 2010).

Wichtig ist auch, dass die Studenten sich aktiv am Lösungsprozess beteiligen anstatt diesen nur nachzuverfolgen. Durch die eigene Denkleistung können sich die oben beschriebenen kognitiven Strukturen erst ausbilden und beispielhafte Fälle als Erfahrungen abgespeichert werden. Darüber hinaus sollten die Studenten explizit dazu angehalten werden, Analogien zu bilden, Konzepte miteinander in Verbindung zu bringen und zu vergleichen. Es soll nicht erst abgewartet werden, dass die Studenten dies von alleine tun (Eva 2005). Neues sollte auf Prinzipien aus alten Fällen zurückgeführt oder mit bestehendem Wissen verknüpft werden.

In Kombination mit einem interaktiven Lernansatz, in dem der Lernende seinen Gedankengang offenlegt, können Lehrer so besser erkennen, wo der Studierende Fehler macht. Diese sollten dem Lerner zurückgespiegelt und korrigiert werden. Durch derartige Interaktion entsteht ein effektives Lernsetting (Kassirer 2010; Levine und Bleakley 2013). An diesem Punkt sind zwei Dinge wichtig: Feedback und Metakognition.

Die Bedeutung von Feedback im *Clinical Reasoning*-Lernprozess wird vielerorts herausgestellt (Norman 2005; Bowen 2006; Kassirer 2010; Pinnock und Welch 2014). Kassirer spricht sich dafür aus, dass Feedback sofort gegeben wird, wenn der Fall gerade bearbeitet wurde und die Informationen noch frisch sind. Geschehen kann dies in Form eines „Case Wrap-up", einer Fallnachbesprechung, in der etwaige Fehler hinterher aufgezeigt und genau besprochen werden – wie es aus dem sogenannten *Debriefing* bei Simulationen bekannt ist (Abatzis und Littlewood 2015).

Aber auch schon während des Lösungsprozesses macht Feedback Sinn. Die *Intermediate Reasoning* genannten Zwischenschritte im Denken auf dem Weg von Datenerfassung bis zur diagnostischen oder therapeutischen Entscheidung bieten bereits wertvolles Material zum Überprüfen, Kommentieren und Korrigieren. *Intermediate Reasoning* findet sich sowohl bei Experten als auch bei Anfängern, ersteren ist dieses jedoch häufig nicht mehr im Einzelnen bewusst. Für die Lernenden macht es Sinn, die Zwischenschritte zu artikulieren. Bevor die einzelnen Schritte nicht klar sind, kann für sie daraus kein schneller Gedankenstrom werden – und das Bewusstmachen der einzelnen Schlüsse ermöglicht deren kritische Überprüfung (Levine und Bleakley 2013). Eine Überprüfung und Korrektur ist wiederum die Aufgabe des Lehrers oder des Lernformates. In einer betreuten Lernsituation kann aus jedem artikulierten Zwischenschritt (sowohl des Anfängers als auch des Experten) ein Positiv- oder ein Negativbeispiel für *Clinical Reasoning* werden. Indem etwa im Denkablauf des Studierenden kognitive Fehler gefunden und anhand der offengelegten Expertenlösung zielführendere Überlegungen gezeigt werden, kann *Intermediate Reasoning* nutzbar gemacht werden. Ein solcher metakognitiver Ansatz könnte zusätzlich über die Lernsituation hinaus Frucht bringen, wenn der angehende Kliniker sich daran gewöhnt, eigene Gedankenschrit-

te zu reflektieren. Ob durch das Lernen von Metakognition tatsächlich zukünftige Praxisfehler im *Clinical Reasoning* reduziert werden, ist noch nicht ausreichend erforscht (Kassirer 2010; Stark et al. 2011).

1.4 Lernen aus Fehlern

Studenten in die Lage zu bringen, eigene Fehler aufzudecken, zu verstehen und zu überdenken, soll das Lernen aus diesen Fehlern ermöglichen. Schließlich stellen Fehler gerade in Simulations- und Übungssituationen eine große Chance dar, Schwachstellen aufzudecken und zu verbessern, die man nicht verstreichen lassen sollte. Gerade spezifisch eigene oder typische Fehler bieten, wenn sie aufgezeigt werden, eine Möglichkeit, das Lehrangebot auf den individuellen Lerner und seine Bedürfnisse abzustimmen (Eva 2005).

In der Medizin ist die Reduktion von Fehlern selbstverständlich von besonders großem Interesse, dadurch dass die Folgen für Gesundheit und Leben von Menschen gravierend sein können. Graber et al. stellten eine Taxonomie von diagnostischen Fehlern in der inneren Medizin auf, in der verschiedene Arten von Fehlern anhand ihrer Ursache unterschieden werden (Graber et al. 2005). Kognitive Fehler, abgegrenzt von Systemfehlern und unverschuldeten Fehldiagnosen, kamen in ihren Untersuchungen in 74 von 100 falsch diagnostizierten Fällen vor, durchschnittlich 4,3 kognitive Fehler pro Fall. Unzureichendes Wissen war dabei die weitaus seltenste Ursache für kognitive Fehler – in den meisten Fällen scheiterten die Ärzte an fehlerhafter Synthese oder Verarbeitung von Informationen, beispielsweise durch Fehleinschätzung der Relevanz von Befunden oder durch Nichtberücksichtigung alternativer Möglichkeiten, sobald eine initiale Verdachtsdiagnose gestellt wurde. Die verfrühte Annahme einer Verdachtsdiagnose – unter Unterlassen weiterer diagnostischer Maßnahmen zum Ausschluss von Differentialdiagnosen – wird *Premature closure* genannt und von Graber et al. (2005) als häufigster Mechanismus kognitiver diagnostischer Fehler herausgestellt.

Die häufigsten Diagnosefehler beruhen nach dieser Analyse also auf fehlerhaftem *Clinical Reasoning*. Es wurden bereits eine Reihe von Strategien vorgeschlagen, um diese Fehler zu reduzieren. Unter dem Stichwort *DeBiasing* werden Strategien zusammengefasst, die darauf abzielen, kognitive Fehler in der Entscheidungsfindung zu entlarven und zu entschärfen. Metakognitive Fähigkeiten zu trainieren ist eine davon (Croskerry 2003; Graber et al. 2005). Information über kognitive Fehler in die Lehre einzubinden, also beispielsweise den Studenten verschiedene Formen von diagnostischem *Bias* aufzuzeigen, wird vorgeschlagen (Croskerry 2000; Hall 2002; Redelmeier 2005) unter der Annahme, dass Wissen über Fehlermechanismen das Bewusstwerden derer in der eigenen Arbeit fördert, um diese zu umgehen oder möglicherweise sogar deren Auftreten zu verhindern. Dagegen argumentieren Norman und Eva, dass die Effektivität solcher Strategien nicht nachgewiesen wurde und im Hinblick auf zugrundeliegende Prinzipien kognitiver Psychologie nicht plausibel sei – da Vorwissen am unbewussten Charakter des kognitiven *Bias* nichts ändere und da nicht ge-

geben sei, dass das Lernen eines Fehlermechanismus themenungebunden sei und damit generell von einem Fall auf den nächsten transferiert werden könne (Norman und Eva 2010). Schließlich wird, wie oben erwähnt, für *Clinical Reasoning* insgesamt eine hohe Kontextspezifität angenommen. Ob Strategien zur Reduktion von fehlerhaftem *Clinical Reasoning* daher als übergreifende, kontextunabhängige Fähigkeiten gesehen werden können, ist fraglich.

Auch hier bietet sich daher zunächst eine fallbezogene Herangehensweise an, um zumindest die häufig begangenen Fehler in üblichen klinischen Situationen im Training abzudecken.

Häufig begangene Fehler in der klinischen Praxis zu identifizieren und zu reduzieren ist auch das Ziel der deutschen Klug Entscheiden-Kampagne (zuerst ausgehend von der Deutschen Gesellschaft für Innere Medizin (DGIM), mittlerweile von mehreren anderen Fachgesellschaften adaptiert), die aus dem Bestreben entstand, durch medizinische Unter- oder Überversorgung verursachte Behandlungskosten und Gesundheitsrisiken für Patienten zu reduzieren (Fölsch et al. 2017). Der übermäßige Gebrauch diagnostischer oder therapeutischer Maßnahmen kann genauso wie deren Unterlassen Resultat fehlerhaften *Clinical Reasoning*s sein. Stattdessen sollen nun spezifische Empfehlungen der Fachgesellschaften in den Abwägungsprozess behandelnder Ärzte einfließen und deren Entscheidungen zugunsten effizienterer Handlungsstrategien beeinflussen. Um die Umsetzung dieser Empfehlungen auf den Weg zu bringen, wurde ein Lehrkonzept entwickelt, mit dem deren Anwendung geübt und geprüft werden kann, und an Medizinstudierenden getestet (Goldmann et al. 2016; Goldmann et al. 2017). Hier zeigte sich, dass viele der Empfehlungen von Medizinstudierenden noch nicht beherrscht wurden – aber auch, dass das genutzte Lernformat geeignet war, um sowohl den Anteil positiver Antworten zu erhöhen, als auch, in vielen Fällen, den Anteil ausgewählter Falschantworten zu reduzieren.

Es handelte sich dabei um ein fallbasiertes Prüfungsformat, das als *Key Feature* bekannt ist. Worum es dabei geht und wieso es zum Lernen von *Clinical Reasoning* geeignet ist, soll in Kapitel 1.6 erläutert werden.

Bevor näher darauf eingegangen wird, ist es jedoch angebracht zu klären, warum ein Prüfungsformat als solches eingesetzt wird, um ein Lernziel zu erreichen. Im folgenden Abschnitt soll daher ein Exkurs über die Bedeutung des Prüfens unternommen werden.

1.5 *Test-Enhanced Learning*: wie Prüfen und Lernen zusammenhängen

Im Fokus dieser Arbeit steht die Aneignung von Kompetenz im klinischen Denken, die Reduktion und Verhinderung von Fehlern im ärztlichen Entscheidungsprozess und all dies vor dem Hintergrund der Fragestellung, wie Medizinstudenten dies am besten lernen können. Prüfungen sind traditionell zwar immer organisatorisch verknüpft mit Lehre, aber als

separate Einheit – mit anderer Zielsetzung, anderer Methodik und anderem Setting. Wieso sich nun aus dem Repertoire der Prüfungen bedient wird, um *Clinical Reasoning* zu lernen, soll aus diesem Abschnitt deutlich werden.

Im bestehenden Bildungssystem sind Prüfungen ein fester Bestandteil des Lernzyklus. Sie dienen der Erfassung des Lernfortschrittes und der Bewertung des Leistungsstandes. Für Lehrer wie Lernende bieten sie Messinstrumente für den Lernerfolg und Maßstab für den Leistungsvergleich. Das Medizinstudium ist dabei keine Ausnahme.

Prüfungen haben jedoch mehr als nur eine Messfunktion, sie nehmen auch direkt und indirekt Einfluss auf die Retention, also das Behalten des Gelernten. Anstatt Wissen lediglich neutral darzustellen und zu bemessen, modifiziert das Testen tatsächlich das Gedächtnis für Wissen und nimmt damit am Lernprozess teil. Diese weitere Funktion des Prüfens hat gegenüber dessen traditioneller Rolle in der Bildung bislang wenig Beachtung gefunden. Dabei ist sie seit langer Zeit bekannt und vielfach beschrieben worden (James 1890; Bacon 1620/2000).

Wird gelerntes Wissen abgeprüft, so wird dieses Wissen besser behalten als wäre es nicht geprüft worden. Über kürzlich angeeignetes Wissen geprüft zu werden fördert die Retention des Gelernten sogar mehr, als die Inhalte ein weiteres Mal einzustudieren. Dieses Phänomen nennt sich „direkter *Testing Effect*". Roediger und Karpicke beschreiben in einer umfassenden Übersichtsarbeit, wie dieser Effekt in vielen Experimenten unter verschiedenen Konditionen repliziert wurde (Roediger und Karpicke 2006b).

Dabei wirkt sich das Testen vor allem auf das Langzeitgedächtnis aus. Beim Vergleich der langfristigen Retention erzielt wiederholtes Testen deutlich bessere Ergebnisse als wiederholtes Lesen des Lernstoffs. Der direkte *Testing Effect* persistiert auch über lange Zeiträume von mehreren Monaten (Nungester und Duchastel 1982). Kurzfristig kann der Vergleich jedoch andersherum ausfallen. Daraus ergibt sich ein Erklärungsansatz für den ebenfalls belegten Umstand, dass Studenten die Effekte von Lesen und Testen auf ihren Lernerfolg mehrheitlich gegenteilig einschätzten: da wiederholtes Ansehen des Lernstoffs kurzfristig höheren Erfolg verspricht als Testen, bevorzugen Studenten die erstere Form des Lernens, obwohl sie im Hinblick auf langfristige Retention ineffektiv ist (Roediger und Karpicke 2006a).

Über den Grund, warum Tests zu einer besseren Retention des Getesteten führen, gibt es verschiedene Theorien. Dass der *Testing Effect* nur auf zusätzliche Exposition gegenüber des Materials zurückzuführen ist, wurde empirisch widerlegt (Roediger und Karpicke 2006a; Raupach et al. 2016). Ein anderer Erklärungsansatz bezieht sich auf die größere Arbeitsleistung, die beim Geprüftwerden gegenüber passiver Exposition mit Lernmaterialien erbracht wird: das aufwändige Hervorholen von Informationen aus dem Gedächtnis (*Retrieval*) wird mit veränderter Bahnung zur besseren Verfügbarkeit der Informationen in Verbindung gebracht. *Retrieval* könnte auch mit verstärkter Elaboration und damit besserer Vernetzung einhergehen, wodurch wiederum Anknüpfpunkte für *Retrieval* vermehrt würden (Roediger

und Karpicke 2006b). Das Resultat kann man so formulieren, dass der Akt des Abfragens von Informationen aus dem Gedächtnis das Gedächtnis für diese Informationen stärkt (Larsen et al. 2008). Weiterhin wird das Konzept der Transfer-gerechten Verarbeitung angeführt, wonach das Testen eine *Retrieval*-Situation produziert, die dem späteren Zurückgreifen auf das Wissen sehr ähnlich sei – es trainiere also die spätere Verwendung des Wissens. Wenn eine Information zu einem späteren Zeitpunkt benötigt würde, sei dies nach dieser Theorie nichts anderes als eine weitere Prüfungssituation. Die kognitiven Bemühungen beim initialen Testen ließen sich also in eine spätere *Retrieval*-Situation transferieren, was diese möglicherweise erleichtere (Roediger und Karpicke 2006b).

Neben den beschriebenen direkten Effekten des Prüfens auf das Lernen – dem eigentlichen *Testing Effect* – gibt es auch indirekte Effekte. Sie werden vermittelt durch das, was Prüfungen in Studenten auslösen: zum Beispiel die Motivation, Zeit und wirksame Methoden zum Lernen zu finden. Effizient ausgedrückt wird dieses Phänomen durch das Axiom *Assessment drives learning* (Wood 2009). Darin liegt auch das Potential von Prüfungen, nicht nur durch ihre Existenz, sondern auch anhand ihres Inhalts und Designs das Lernverhalten von Prüflingen zu steuern (Newble und Jaeger 1983). Durch häufigere Tests beschäftigen die Studenten sich in der Vorbereitung häufiger und intensiver mit dem Lernstoff, was zu besseren Lernergebnissen führt. Abgesehen vom quasi erzwungenen Lernen durch Klausurvorbereitung können Studenten auch von Tests profitieren, indem sie ihre Schwächen aufdecken und ihre Lernbemühungen danach gezielter steuern (Larsen et al. 2008). Ebenso können indirekte Effekte auf das Lernen durch das Lehrpersonal vermittelt werden, wenn diese beispielsweise anhand der Prüfungsergebnisse ihre Lehre anpassen.

Die indirekten Effekte des Testens scheinen naheliegender zu sein und haben häufig mehr Beachtung gefunden als der direkte *Testing Effect*, selbst dann, wenn indirekte Effekte eine unzureichende Erklärung für die Überlegenheit häufigen Testens lieferten (Fitch et al. 1951). Das kann daran liegen, dass der Nachweis des direkten Effekts außerhalb kontrollierter Laborbedingungen schwieriger ist, jedoch liefern auch angewandte Studien Evidenz für die Generalisierbarkeit des *Testing Effect*s (Roediger und Karpicke 2006b). Dabei gehen die indirekten und die direkten Effekte des Testens sicherlich Hand in Hand, was letztendlich für die Lerner nur förderlich ist.

Um vom *Testing Effect* zu profitieren, macht es also Sinn, Prüfungen in den Lernprozess einzubauen. Gerade im Medizinstudium ist es angesichts der großen Menge an Lernstoff – und seiner gewichtigen Relevanz – erstrebenswert, nicht nur kurzfristigen Lernerfolg, sondern langfristige sichere Retention des Gelernten zu erlangen. Als *Test-Enhanced Learning* (TEL) wird der Ansatz bezeichnet, Lernen und Gedächtnis durch gezielten Einsatz von Tests im Bildungskontext zu verbessern. Larsen et al. empfehlen den Einsatz und die weitere Erforschung von *Test-Enhanced Learning* im Medizinstudium (Larsen et al. 2008), da es erwiesenermaßen die Retention von Faktenwissen fördere und da klinische Problemlösekompetenz auf einer soliden Wissensbasis beruhe. Aber nicht nur Faktenwissen kann mit-

tels *Test-Enhanced Learning* verbessert werden. In Studien wurde eine Überlegenheit des Testens auch für den Erwerb praktischer Fertigkeiten (Kromann et al. 2009) und höherer kognitiver Funktionen wie kritischem Denken (Dobson et al. 2018) in Bezug auf Inhalte des Medizinstudiums belegt. Eine Übertragung der vielversprechenden Erkenntnisse über *Test-Enhanced Learning* auf die besonderen Lernziele der medizinischen Ausbildung konnte also erfolgen. In den vergangenen Jahren wurde die Evidenz für die Effektivität von *Test-Enhanced Learning* im Medizinstudium stetig erweitert und aufgrund guter Ergebnisse dessen curriculare Implementierung empfohlen (Green et al. 2018; Raupach und Schuelper 2018).

Wie die beste Umsetzung von *Test-Enhanced Learning* im Medizinstudium aussieht, lässt noch Raum für Diskussion, jedoch lassen sich ein paar evidenzbasierte Empfehlungen aus der Grundlagenforschung zum *Testing Effect* ziehen:

Häufigeres, regelmäßiges Testen wirkt sich besser auf den Lernerfolg aus als einmaliges Testen (Roediger und Karpicke 2006b). Prüfungen sollten also wiederholt über den Lernzeitraum hinweg stattfinden (Larsen et al. 2008). Der *Testing Effect* wurde für verschiedene Prüfungsformate nachgewiesen, auch für die häufig im Medizinstudium genutzten *Multiple Choice*-Fragen (MCQs) (Nungester und Duchastel 1982). Unabhängig davon, welches Fragenformat untersucht wurde, Geprüftwerden führte immer zu besserer längerfristiger Retention als alleiniges Studieren des Lernmaterials. Im Vergleich von verschiedenen Formaten zeigte sich jedoch, dass Tests, die die Produktion von Antworten erforderten (zum Beispiel frei formulierte Kurzantworten) einen größeren Lerneffekt erzielen als solche, bei der die richtige Antwort nur wiedererkannt werden muss (zum Beispiel *Multiple Choice*-Tests; Butler und Roediger 2007; McDaniel et al. 2007). Dies stimmt überein mit der Erkenntnis, dass die eigene Produktion von Material während des Lernens mehr Retention bewirkt als alleiniges Lesen des Materials – dem sogenannten *Generation Effect* (Roediger und Karpicke 2006b). Zudem wird argumentiert, dass die Antwortproduktion mit schwierigerem *Retrieval* einhergeht als das Wiedererkennen von Antworten und mühsameres *Retrieval* eine bessere Retention zur Folge hat (Larsen et al. 2008).

Weiter erhöht wird die Effektivität des Testens durch die Gabe von Feedback. Wichtig ist, dass der *Testing Effect* auch ohne jegliches Feedback in Erscheinung tritt, aber durch Feedback kann er zusätzlich verstärkt werden. Darin sollte die richtige Antwort offenbart werden, unabhängig vom sonstigen Inhalt des Feedbacks (Larsen et al. 2008).

Zusammengefasst lässt sich daraus schließen, dass wiederholtes Abprüfen von Wissen oder Kompetenzen, möglichst mit offenen Fragen und Feedback über die korrekte Antwort, Vorteile für den Lernerfolg im Medizinstudium verspricht.

Um diese zu realisieren, bietet sich die Einrichtung von formativen Prüfungen an. Im Gegensatz zu summativen Prüfungen, die der Benotung des Leistungsstands dienen, werden formative Prüfungen zur Unterstützung des Lernprozesses eingesetzt. Die beiden Arten von Prüfungen unterscheiden sich darin, wozu ihre Ergebnisse produziert und genutzt werden: in summativen Prüfungen zur Bewertung der Prüflinge, in formativen Prüfungen

zur Verbesserung der Lehre und des Lernens durch direkte und indirekte Effekte des Testens. Roediger und Karpicke führen noch Evidenz an, dass formative Prüfungen das Lernen verbessern und bringen es mit ihrer Formulierung auf den Punkt: „formative assessment is often referred to as assessment *for* learning, in contrast to assessment *of* learning." (Roediger und Karpicke 2006b).

1.6 *Key Feature*-Fragen zur Prüfung von *Clinical Reasoning*

Nachdem die Bedeutung von Prüfungen für das Lernen, auch von komplexen Fähigkeiten, herausgestellt wurde, soll es nun um ein geeignetes Prüfungsformat für die Evaluation von *Clinical Reasoning* gehen.

Die hohen Ansprüche an praktizierende Ärzte bedingen die Notwendigkeit verlässlicher Überprüfung, ob Kandidaten die Voraussetzungen für diese Tätigkeit erfüllen. Wie oben beschrieben gehört neben der breiten und tiefen Wissensbasis, die im Studium vermittelt wird, klinisches Denken als Kernkompetenz zu diesen Voraussetzungen für erfolgreiches ärztliches Handeln. Ebenfalls wurde festgestellt, dass *Clinical Reasoning* sehr komplex ist und nur teilweise verstanden wird. Eine geeignete Prüfungsmethode dafür zu finden, stellt daher ein anhaltendes Problem dar, an welchem aber gleichzeitig ein großes Interesse besteht. Daher wurden bereits viele Bemühungen der Erforschung von Prüfungsformen für klinische Entscheidungskompetenz unternommen. Herausgehoben werden sollen hier *Key Feature*-Prüfungen, die im Gegensatz zu anderen in der Vergangenheit entwickelten Ansätzen markante Vorteile aufweisen, unter anderem in der Anwendung und Auswertung, aber auch in ihrer Validität durch ihre Art, verschiedene Aspekte von *Clinical Reasoning* aufzugreifen (siehe Farmer und Page 2005; Hrynchak et al. 2014). Nach einer kurzen Vorstellung des *Key Feature*-Konzeptes sollen diese Vorteile hier erläutert werden.

Als *Key Feature*s werden spezifische Schlüsselstellen in einem Fall bezeichnet, solche inhaltlichen und zeitlichen Punkte, an denen wegweisende Entscheidungen getroffen werden müssen. Die Idee von *Key Feature*s entstammt der medizinischen Ausbildung und bezieht sich daher von vorneherein auf klinische Patientenfälle (auch wenn eine Übertragung des Konzeptes in andere Bereiche prinzipiell denkbar ist). Ein *Key Feature* ist also ein Schritt in der Handhabung eines Patientenfalls, der diesen Fall maßgeblich definiert und prägt, sowohl in Bezug auf seine Interpretation als auch auf seinen Ausgang. Es bezeichnet ein Kernelement in der klinischen Problemlösung. Wird eines dieser Elemente verändert, etwa bei einer Fallsimulation, verändert sich der ganze Fall und mit ihm die Fähigkeiten und Strategien, die zur Lösung erforderlich sind (Page et al. 1995). Damit einhergehend sind die *Key Feature*s die Stellen der größten Herausforderung für den Entscheidung tragenden Arzt in diesem Patientenfall (Bordage und Page 2018). *Key Feature*s werden daher ebenfalls beschrieben als schwierige Aspekte in einem Fall, die Punkte, an denen Studenten am häufigsten Fehler machen (Hrynchak et al. 2014).

Die Entstehung des *Key Feature*-Ansatzes kann in Verbindung gesetzt werden mit der Entdeckung des Problems der Kontextspezifität in der Erforschung von klinischem Denken. Nachdem die Suche nach einer allgemeinen medizinischen Problemlösefähigkeit erfolglos blieb und stattdessen gezeigt wurde, dass klinische Entscheidungskompetenz in hohem Maße fallspezifisch ist (Elstein et al. 1978), verschob sich der Fokus auf die spezifischen inhaltlichen Elemente von Patientenfällen, die Angriffspunkte für eine validere Messung klinischen Denkens darstellen sollten. Auf der 1984 stattfindenden *Cambridge Conference on the Assessment of Medical Competence* wurde festgestellt, dass Kontextspezifität daher rühre, dass die Lösung jedes klinischen Problems auf der Lösung weniger Kernelemente beruhe (Fragen aus den Bereichen Diagnostik und Therapie), welche für den Fall meist einzigartig seien (Norman et al. 1984). Bordage und Page übertrugen diese Erkenntnis ins Prüfungsdesign (Bordage und Page 1987): Wenn die erfolgreiche Lösung klinischer Probleme tatsächlich an spezifischen Kernfragen hängt, sollte Kompetenz anhand der Beantwortung dieser Fragen – was in der Praxis dem Umgang mit diesen spezifischen Herausforderungen entspräche – feststellbar sein. Die damit gemessene Kompetenz ist zunächst fallspezifisch. Bezieht sich die Prüfung jeweils nur auf die entscheidendsten und schwierigsten Aspekte in einem Fall, wird die Bearbeitung auf diese wenigen Fragen verkürzt und somit können viele Fälle in kürzerer Zeit abgeprüft werden. Mit hohen Fallzahlen soll eine bessere Repräsentation der inhaltlichen Breite klinischer Situationen erreicht und damit dem Problem der Kontextspezifität begegnet werden.

Ein *Key Feature*-Fall besteht aus einer Beschreibung des Szenarios – der sogenannten Fallvignette – und den daran anknüpfenden Fragen, die sich auf die *Key Feature*s des Falles beziehen. Diese werden gemäß ihrer Funktion in der Prüfung als Items bezeichnet. In der Regel gehört zu jedem *Key Feature* ein Item. Das heißt, jedes *Key Feature* wird als Frage formuliert und der Prüfling muss mit seiner Antwort die zu treffende Entscheidung darstellen. Die Anzahl an *Key Feature*s in einem Fall ist beliebig.

In früheren Ansätzen zur Prüfung von klinischem Denken wurde auf zeitaufwändige, minutiös ausgearbeitete Patientenfälle gesetzt, für deren Bearbeitung eine Repräsentativität für *Clinical Reasoning*-Kompetenz angenommen wurde, beispielsweise die sogenannten *Patient Management Problems* (PMPs). Für diese wurde jedoch eine unzureichende Reliabilität bei geringer Korrelation zwischen den Ergebnissen verschiedener PMPs gezeigt (Page und Bordage 1995), was im Nachhinein als weitere Bestätigung der zuvor von Elstein et al. beschriebenen Kontextspezifität auffällt. Ihr Einsatz zur Prüfung klinischer Entscheidungskompetenz wurde wieder aufgegeben. Anstatt einen generellen Entscheidungsprozess anzunehmen und diesen, wie im Falle der PMPs, anhand weniger langwieriger Fallbearbeitungen zu prüfen, wird mit dem *Key Feature*-Format anhand vieler kleiner Einzelfälle ein Portfolio aus klinischen Situationen und Wissensbereichen abgefragt. Das Potential, Kompetenz in klinischem Denken abzubilden, müsste bei *Key Feature*-Prüfungen nach dem aktuellen Stand des Wissens also größer sein.

1.6.1 Eigenschaften von *Key Feature*-Prüfungen

Tatsächlich gibt es guten Grund zur Annahme, dass *Key Feature*-Prüfungen für die Evaluation von *Clinical Reasoning* geeignet sind. Testungen der Validität und Reliabilität haben in der Vergangenheit erfolgversprechende Ergebnisse geliefert (Hrynchak et al. 2014; Bordage und Page 2018). In verschiedenen Studien über *Key Feature*-Prüfungen konnte insbesondere Konstruktvalidität für *Clinical Reasoning* gezeigt werden. Beispielsweise konnten *Key Feature*-Prüfungen die Auswirkungen von Lehrinterventionen abbilden, die darauf abzielten, *Clinical Reasoning*-Kompetenz zu verbessern. Auch größere klinische Expertise und Erfahrung stellte sich in vielen Studien durch bessere Ergebnisse in entsprechenden *Key Feature*-Prüfungen dar. Dass anhand von *Key Feature*-Prüfungen zwischen unterschiedlichen Trainingslevels in klinischer Medizin unterschieden werden kann, wurde wiederholt berichtet, was passend ist zu der Annahme, dass *Clinical Reasoning* maßgeblich durch die Exposition gegenüber Patientenfällen trainiert wird. Auch Studien, die Denkprozesse bei *Key Feature*-Fragen im Vergleich zu *Multiple Choice*- oder anderen Fragen verglichen, fielen zugunsten der These aus, dass bei ersteren komplexere kognitive Vorgänge im Sinne von klinischem Denken aktiviert werden. Korrelationen von Ergebnissen in *Key Feature*-Prüfungen und späterem Erfolg in klinischer Praxis – wenn auch nur an wenigen sehr spezifischen Endpunkten gemessen – weisen auf eine prognostische Validität dieser Prüfungen hin (Hrynchak et al. 2014; Bordage und Page 2018).

Obwohl alle Bemühungen, Validität für ein so komplexes und mitunter unscharfes Konstrukt wie *Clinical Reasoning* nachzuweisen, als Annäherungsversuche gesehen werden müssen (Bordage und Page merken an, dass nicht das eigentliche Denken des Klinikers gemessen wird, sondern dessen Ergebnis in Form der sich daraus ergebenden Entscheidungen und Handlungen (Bordage und Page 2018)), spricht die bisher gesammelte Evidenz in der Summe dafür, *Key Feature*-Prüfungen eine akzeptable Validität für dieses Gebiet zuzuschreiben; selbstverständlich unter der Voraussetzung, dass sie gut gemacht sind.

Ob das der Fall ist, lässt sich unter anderem anhand der gemessenen Reliabilität feststellen. Je höher die Reliabilität einer Prüfung, desto zuverlässiger sind die Ergebnisse unter denselben Bedingungen reproduzierbar. Empirischen Daten zufolge wird die Reliabilität einer *Key Feature*-Prüfung unter anderem von der Teststruktur, dem Antwortformat und sprachlichen Aspekten beeinflusst (Hrynchak et al. 2014). Um zuverlässige Ergebnisse bezüglich der *Clinical Reasoning*-Kompetenz des Prüflings zu erhalten, muss die Zahl an *Key Feature*-Fällen hoch genug sein. Sie sollte im Bereich von 25-40 (Hrynchak et al. 2014) bzw. >35-40 (Bordage und Page 2018) Fällen pro Prüfung liegen. Selbstverständlich müssen die Fälle sorgfältig ausgewählt werden, ebenso wie die *Key Feature*s innerhalb eines Falles: sie sollten eine möglichst hohe Trennschärfe aufweisen (die Fähigkeit, zwischen guten und schlechten Prüflingen zu unterscheiden). Die Nutzung von *Key Feature*-Fällen und –Items mit hoher Trennschärfe – die wiederum im Vorfeld empirisch erhoben werden muss – sorgt für eine höhere Reliabilität der Prüfung (Bordage und Page 2018). Variabel ist an der Teststruktur

auch die Anzahl der *Key Feature*s, und damit der Fragen, pro Fall. Laut einer Untersuchung anhand von kanadischen Prüfungsergebnissen soll bei einer Anzahl von 2-3 *Key Feature*s pro Fall die höchste Reliabilität erreicht werden – während ein einziges *Key Feature*-Item zu wenig Informationen liefere, trügen mehr als drei Items pro Fall nicht mehr zu einer Erhöhung der Reliabilität bei (Norman et al. 2006). Außerdem wurde gezeigt, dass Laiensprache in den Beschreibungen der Fallszenarien – im Gegensatz zu medizinischen Fachtermini – zu einer besseren Trennschärfe führt (Eva und Wood 2003; Eva et al. 2010). Damit werden nicht nur die Prüfungsergebnisse in ihrer Aussagekraft verlässlicher, sondern diese Tatsache liefert auch ein weiteres Argument für die Konstruktvalidität für klinische Expertise, da größere Vertrautheit im Umgang mit Laiensprache auf mehr klinische Erfahrung schließen lässt und davon ausgegangen werden kann, dass Expertise mit größerer klinischer Erfahrung einhergeht (Hrynchak et al. 2014).

In *Key Feature*-Prüfungen können verschiedene Fragenformate zum Einsatz kommen. Grundsätzlich gibt es offene und geschlossene Fragen: in der offenen Variante muss die Antwort frei und eigenhändig eingetragen werden, bei geschlossenen Formaten werden Antwortmöglichkeiten vorgegeben. Die Länge der erwarteten Antworten definiert verschiedene Arten von Freitextfragen, von kurzen Schlagwörtern bis hin zu Aufsätzen. Bei geschlossenen Fragenformaten ist die Menge der Antwortmöglichkeiten variabel, sowie die Menge der korrekten Antworten darunter. Relevante Antwortformate, zu denen Daten über ihre Eignung für *Key Feature*-Prüfungen vorliegen, werden im folgenden Abschnitt kurz erklärt.

Eine höhere Reliabilität erreichen *Key Feature*-Prüfungen, wenn offene Fragen gestellt werden – freie Antworten eignen sich besonders gut zur Diskriminierung im leistungsschwächeren Bereich (Bordage und Page 2018). Freie Antwortformate erfordern die selbständige Produktion der Antwort ohne Darbietung von Antwortmöglichkeiten. Aber auch der Einsatz von *Short Menu*-Fragen, bei denen die Antworten aus einer Liste von durchschnittlich 15 bis 20 Möglichkeiten ausgewählt werden, hat laut den kanadischen Autoren seine Berechtigung: Die Prüfungsergebnisse fielen mit *Short Menu*-Fragen durchschnittlich 20% höher aus als bei freien *Write-In*-Antworten, jedoch war die Reliabilität gemessen an den Testergebnissen bei Short-Menu-Klausuren dieselbe wie bei solchen Prüfungen, in denen das Short-Menu- und das *Write-In*-Format gemischt vorkamen (Page und Bordage 1995).

Aber auch mit anderen Antwortformaten konnten gute Ergebnisse erzielt werden: Das sogenannte *Long Menu*-Format wird insbesondere in computerbasierten *Key Feature*-Prüfungen angewandt (Schuwirth et al. 1996). Dabei soll der Prüfling seine Antwort aus einer Liste von mehreren hundert Antwortmöglichkeiten auswählen. In der computergestützten Variante wird er dazu angehalten, seine eigene Antwort in ein Textfeld einzugeben, woraufhin sich ein *Drop Down*-Menü öffnet, das alle passenden hinterlegten Antworten (mit der gleichen Buchstabenfolge) anzeigt. Bis zur Eingabe einer Mindestanzahl an Buchstaben sind keine Antwortmöglichkeiten zu sehen und erst durch die Eingabe eines spezifischen

Begriffes werden die diesen Begriff beinhaltenden Optionen sichtbar. Damit stellen *Long Menu*-Fragen eine Art Mittelweg zwischen offenen und geschlossenen Antwortformaten dar. Über das *Long Menu*-Fragenformat wird gesagt, dass es im Vergleich zum *Multiple Choice*-Format *Cueing*-Effekte reduziert, also ein Hinweisen auf die richtige Antwort durch ihre Sichtbarkeit in einer Antwortliste (Schuwirth et al. 1996; Rotthoff et al. 2006; Cerutti et al. 2016). Da die Antwort bei diesem Fragenformat erst selbstständig erdacht werden muss, bevor sie ins Textfeld eingegeben und im *Drop Down*-Menü angezeigt wird, ähnelt das Format dem freien Antworten. Da nur Antworten gegeben werden können, die in der Antwortliste hinterlegt sind, muss die richtige Antwort jedoch am Ende aus den passenden Optionen im *Drop Down*-Menü ausgewählt werden. Sich dadurch ergebende Wiedererkennungseffekte sollen durch eine möglichst große Anzahl an hinterlegten Antworten und Falschantworten reduziert werden. Eine weitere Konsequenz für die Prüfenden ist, dass eine zuweilen recht große Zahl richtiger Antworten hinterlegt werden muss. Unter Umständen werden auch bei der Betrachtung der studentischen Eingaben auch im Nachgang einer Prüfung noch Antworten identifiziert, die als korrekt zu werten sind.

Tests, die eine eigene Antwortproduktion im Gegensatz zur Wiedererkennung der richtigen Antwort erfordern, erzielen im Sinne des direkten *Testing Effect* eine bessere Retention des Wissens (Larsen et al. 2008). Das *Long Menu*-Format hat eine starke Antwortproduktions-Komponente, eine gewisse Wiedererkennungs-Komponente kann jedoch nicht ausgeschlossen werden. Einen klaren Vorteil bieten *Long Menu*-Fragen in der erleichterten Auswertung gegenüber Freitext-Fragen, die die begrenzte Anzahl an Antwortmöglichkeiten mit sich bringt. Außerdem wird die Gefahr einer Konstrukt-irrelevanten Varianz der Ergebnisse durch orthographische Fehler beim Eintragen einer Freitext-Antwort vermieden. Bei ähnlichen Ergebnissen, wie offene Fragen sie erbringen, bieten *Long Menu*-Fragen daher einen „akzeptablen Ersatz“ für Freiantworten (Schuwirth et al. 1996). Sinnvollerweise sollten *Long Menu*-Fragen nur dann eingesetzt werden, wenn kurze, präzise und klar formulierbare Antworten gegeben werden können – dann entsprächen sie am ehesten dem freien *Short Answer*-Format (Rotthoff et al. 2006).

Long Menu-Fragen haben, neben *Short Menu*- und freien Antwortformaten, im Rahmen von *Key Feature*-Prüfungen in der medizinischen Ausbildung einen festen Platz gefunden (Hrynchak et al. 2014).

Der gemischte Einsatz von offenen und geschlossenen Antwortformaten wird auch mit der Begründung unterstützt, dass die Fragen dadurch verschiedene Situationen aus der tatsächlichen Praxis authentischer abbilden könnten: Da einige Entscheidungen im Management des Patienten mit nahezu unbegrenzten Optionen einhergingen, während andere auf eine definierte Anzahl an Alternativen beschränkt seien – beispielsweise im Rahmen der Labordiagnostik – mache es Sinn, das Antwortformat an die in der Frage repräsentierte Situation anzupassen (Schuwirth 1998). Hierin liegt ein entscheidender Vorteil von *Key Feature*-Prüfungen: Sie sind nicht an ein starres Format gebunden. *Key Feature* ist letztlich kein Prü-

fungsformat an sich, sondern ein Ansatz, klinische Entscheidungen zu prüfen (Bordage und Page 2018). Seine große Flexibilität, verschiedensten klinischen Fragestellungen gerecht zu werden, bedingt, dass nicht der Inhalt an die Prüfung angepasst und damit das geprüfte Konstrukt verändert werden muss, sondern andersherum die Prüfung auf den Inhalt zugeschnitten werden kann.

1.6.2 Computerbasierte *Key Feature*-Prüfungen

Flexibilität bietet der *Key Feature*-Ansatz auch hinsichtlich des Präsentationsformates der Prüfungen. Es können sowohl Papierklausuren als auch computerbasierte *Key Feature*-Prüfungen geschrieben werden. Letztere bieten einige Vorteile:

Bild- und Videomaterial kann in elektronischer Form unkompliziert und in hoher Qualität eingebunden werden. Damit kann die Bandbreite an abgefragten Kompetenzen erweitert und die Realitätsnähe des Prüfungsfalls verbessert werden. Außerdem kann das Vor- und Zurückspringen innerhalb eines Falls verhindert werden, sodass zusätzliche Informationen zwischen den einzelnen *Key Feature*-Fragen gegeben werden können, ohne dass dadurch Antworten verraten oder erleichtert werden (Farmer und Page 2005). Der initiale Aufwand bei der Erstellung der Prüfung, der bei *Key Feature*-Prüfungen durch die Notwendigkeit, Fälle zu redigieren, generell hoch ist, kann bei der elektronischen Form höher und ressourcenintensiver ausfallen. Jedoch bietet sie danach eine flexible, am Computer allzeit verfügbare und beliebig oft durchführbare Prüfungsoption. Die Auswertung, wie im vorherigen Absatz bereits erwähnt, ist besonders beim Einsatz von antwortlistenbasierten Fragenformaten unkompliziert und kann automatisch erfolgen.

Die Vorteile des *Long Menu*-Formates können bei computerbasierten *Key Feature*-Prüfungen voll ausgenutzt werden. Ein Antwortkatalog von mehreren tausend Antworten ist kein Problem. Die Handhabung wird durch die elektronische Auswahl der Antworten im *Drop Down*-Menü gegenüber der papierbasierten Variante erheblich erleichtert.

Die Akzeptanz von computerbasierten *Key Feature*-Prüfungen unter Medizinstudenten erwies sich in Studien als mittelmäßig bis gut. Tendenziell bewerteten die Befragten die *Key Feature*-Fälle selbst als positiv (Fischer et al. 2005; Huwendiek et al. 2017), lediglich der Einsatz des Computers wurde in einer 17 Jahre zurückliegenden Studie mit Skepsis betrachtet (Fischer et al. 2005).

1.6.3 Einsatz von *Key Feature*-Prüfungen

Bisher wurden computergestützte *Key Feature*-Prüfungen an verschiedenen Universitäten und von medizinischen Fachgesellschaften weltweit zur Evaluation von klinischem Denken eingesetzt (Hatala und Norman 2002; Fischer et al. 2005; Nikendei et al. 2009; Sullivan et al. 2020) – zu Studienzwecken oder in der curricularen Routine, computer- oder papierbasiert, mit *Long Menu*- und anderen Fragenformaten. In den USA wurde vom *American College of Surgeons* für angehende Chirurgen zu Beginn ihrer Facharztweiterbildung eine *Key Feature*-

Prüfung über klinische Entscheidungsfindung entwickelt (Sullivan et al. 2020), in Kanada werden *Key Feature*-Fälle seit 1992 im staatlichen Abschlussexamen des Medizinstudiums verwendet (Page und Bordage 1995). Der guten Datenlage entsprechend haben *Key Feature*-Prüfungen sich als valides Messinstrument für *Clinical Reasoning* etabliert.

In ihrem Review über die aktuelle Validitätslage zum *Key Feature*-Ansatz rufen Bordage und Page dazu auf, die formative Funktion von *Key Feature*-Prüfungen – speziell in der Kombination formativer und summativer Zwecke – weiter zu erforschen. Der *Key Feature*-Ansatz bietet eine Gelegenheit, Prüflinge mit detaillierten Informationen über ihre klinischen Entscheidungen auszustatten. Diese sollten besser genutzt werden, um eigene Stärken und Schwächen zu identifizieren, sowohl in Bezug auf spezifische klinische Situationen und Disziplinen, als auch im Hinblick auf potentiell schädliche Entscheidungen, die sie im Rahmen einer *Key Feature*-Prüfung getroffen hätten (Bordage und Page 2018).

An der Universitätsmedizin Göttingen werden seit 2013 formative *Key Feature*-Prüfungen im Rahmen der curricularen Lehre eingesetzt. Dass die Vorteile formativen Testens auch mit *Key Feature*-Prüfungen mit einem so komplexen Lernziel wie *Clinical Reasoning* zu erzielen sind, wurde experimentell nachgewiesen (Raupach et al. 2016). Auf dieser Grundlage sind formative *Key Feature*-Prüfungen fester Bestandteil der Lehre auf dem Gebiet der inneren Medizin im klinischen Studienabschnitt geworden. Des Weiteren wurde auch der Einsatz von Video-*Feature*s erprobt (Ludwig et al. 2018).

Trotz aller theoretischer und praktischer Vorteile und der über Jahre hinweg gesammelten Evidenz, die gute Argumente für den Einsatz von TEL und *Key Feature*-Prüfungen zum Lernen von *Clinical Reasoning* liefern, konnten in vielen Studien – trotz messbarer signifikanter Verbesserung des Lernerfolgs – nur mittelmäßige Ergebnisse hinsichtlich der durchschnittlichen Gesamtpunktzahl in einer Prüfung erzielt werden (Raupach und Schuelper 2018), was sich auch in den an der UMG durchgeführten Studien bestätigte (Raupach et al. 2016; Ludwig et al. 2018). Eine mögliche Erklärung dafür ist der formative Charakter der Prüfungen.

1.7 Weiterentwicklung des *Key Feature*-Formates durch Kontrastierung mit Falschantworten

Das Ziel des Einsatzes formativer *Key Feature*-Prüfungen im klinischen Studienabschnitt an der Universitätsmedizin Göttingen ist die Förderung von differentialdiagnostischer und –therapeutischer Entscheidungskompetenz. Dass *Clinical Reasoning* bereits im Medizinstudium trainiert werden muss und auch Fehlüberzeugungen, Verständnisprobleme und Flüchtigkeitsfehler dort bereits direkt adressiert werden können und sollten, wird empfohlen (Norman 2005; Pinnock und Welch 2014). Dass Medizinstudenten im Rahmen ihrer Ausbildung genug Praxiserfahrung sammeln können, um alle Facetten klinischer Kompetenz aufzubauen, geschweige denn alle wichtigen Fehler machen oder mitbekommen können, ist

allerdings unrealistisch (Norman 2012). Hier kommt sorgfältig gestalteter Lehre umso mehr Bedeutung zu.

Im Abschnitt „Lernen aus Fehlern" wurde dargelegt, dass sich für die Reduktion von häufigen Fehlern im klinischen Denken ein fallbasierter, die Kontextspezifität berücksichtigender Ansatz eignen dürfte. Der *Key Feature*-Ansatz wurde genau auf diesem Prinzip basierend entwickelt und die *Key Feature*s selbst sollen die Stellen sein, an denen häufig Fehler gemacht werden. Diese Fehler nicht nur indirekt durch das Lernen der richtigen Antwort anzugehen, sondern auch direkt in den Fokus zu nehmen, um von ihnen zu lernen, scheint nur sinnvoll auf Grundlage dessen, was wir über Fehler wissen: sie bieten einen Einblick in die Denkweise des Lernenden und damit einen idealen Anknüpfungspunkt für Weiterentwicklung (Weingardt 2014). Für die Entwicklung von Expertise auf Gebieten von prozeduralem Wissen ist „Fehlererfahrung" ein förderlicher Baustein (Gartmeier et al. 2008). Gerade in der medizinischen Ausbildung sollte dem fehlerhaften Denken der Studierenden mehr Raum gegeben werden, um dieses erkennen und korrigieren zu können (Levine und Bleakley 2013). Da *Key Feature*-Prüfungen es schaffen, klinisches Denken abzubilden, bieten sie den idealen Rahmen dafür, Fehler zutage zu bringen und alternative Strategien anzuwenden. Außerdem eignet sich das *Key Feature*-Format dazu, Feedback einzubinden. Aus der Literatur über *Test-Enhanced Learning* wissen wir, dass Feedback den Lerneffekt des Testens verstärkt (Larsen et al. 2008). Gerade, wenn aus Fehlern gelernt werden soll, ist Feedback wichtig, um tieferes Verständnis zu ermöglichen und Verwirrung abzuwenden (Stark et al. 2011). *Key Feature*-Prüfungen bieten außerdem den Vorteil, dass eine große Zahl von Studenten damit erreicht werden kann bei minimalem Einsatz zeitlicher und personeller Ressourcen, welche für ein *Coaching*-Setting, wie es gerne für Entwicklung von *Clinical Reasoning*-Kompetenz vorgeschlagen wird (u.a. Kassirer 2010), meist ein limitierender Faktor sind.

Das Bestreben, klinisches Denken und Entscheiden mithilfe von *Key Feature*-Prüfungen zu lernen und dieses Lernen durch die Auseinandersetzung mit Fehlern effektiver zu machen, verspricht also Erfolg. Die mehrjährige Erfahrung mit dem Einsatz formativer *Key Feature*-Prüfungen in der klinischen Lehre an der UMG bietet gute Voraussetzungen dafür: Testinfrastruktur sowie mehrfach getestete und verfeinerte *Key Feature*-Fälle sind vorhanden. Außerdem bergen die über mehrere Semester hinweg gesammelten Prüfungsdaten wertvolle Informationen über häufige Fehler und Schwierigkeiten teilnehmender Studenten Eine systematische Analyse der häufigsten Falschantworten zeigte gehäuft auftretende Fehlüberzeugungen auf (Goldmann et al. 2017). Diese Informationen nutzbar zu machen und zur Intensivierung des Lernens zur Verfügung zu stellen, erwuchs als Idee aus dem hier ausgeführten Wissensstand über das Lernen von *Clinical Reasoning* mit dem Ziel der Weiterentwicklung eines bereits gut funktionierenden Ansatzes. Schließlich lassen die bisher meist durchschnittlichen Prüfungsleistungen der Studierenden, wie oben erwähnt, Raum für Verbesserung.

Stark et al. beobachteten in einer Studie zu fallbasiertem Lernen im Medizinstudium, dass die Konfrontation mit Fehlern kognitiv anspruchsvoller war als die Arbeit mit korrekten Beispielen. Die höhere Schwierigkeit konnte mit elaboriertem Feedback kompensiert werden (Stark et al. 2011). Im Fazit zu ihrer Studie fordern Stark et al. dazu auf, Ansätze zu testen, in denen Studenten mit diagnostischen Fehlern und deren Erklärung herausgefordert werden. Aus der Theorie des *Test-Enhanced Learning* geht hervor, dass kognitiv mühsamere Aufgaben zu verbesserter Lernleistung führen können (Larsen et al. 2008). Studierende im Rahmen von formativen *Key Feature*-Prüfungen mit Falschantworten zu konfrontieren, könnte demnach zu besserer Leistung führen. Schließlich löst die Konfrontation mit Fehlern ein Hinterfragen des eigenen Verständnisses und sogenannte *Self-explanation*-Prozesse aus (inneres Erklären oder Herleiten) und bietet damit die Chance, sich tiefer mit dem Inhalt auseinanderzusetzen. Dies sollte zu tieferem Lernen und Verständnis führen, außerdem zur Vermeidung von falschem Lernen (VanLehn 1999). Die Erklärung von Fehlern soll dem Lernenden ermöglichen, eigene Lösungswege zu reflektieren. Reflektion wird zudem im medizinischen Kontext eine wichtige Rolle bei der Entwicklung von Expertise und tieferem Lernen zugeschrieben. Als Lernstrategie genutzt könne Reflektion bei der Integration von neuen Informationen mit existierendem Wissen und Fähigkeiten helfen (Mann et al. 2009). Die Kontrastierung von Differentialdiagnosen und die explizite Begründung ihrer Priorisierung kann zudem den Aufbau von *Illness Scripts* fördern (Bowen 2006).

Aufgrund der bereits vorhandenen Informationen über häufige Fehler in früheren *Key Feature*-Prüfungen konnte eine Implementierung von fehlerhaften Beispielen in Form von Falschantworten in die formativen Prüfungen leicht erfolgen. Daraus entwickelte sich das in dieser Arbeit untersuchte Lernformat (siehe dazu „Material und Methoden“).

1.8 Fragestellung und Hypothesen

Die hier vorgestellte Studie wurde entwickelt, um folgende Forschungsfragen zu beantworten:

1) Wie wirkt sich die Aufforderung zur Elaboration und Kontrastierung von Falschantworten im Rahmen formativer *Key Feature*-Prüfungen im zusätzlich zur Darbietung eines elaborierten Feedback-Texts auf den Erwerb und die langfristige Retention von *Clinical Reasoning* aus?
2) Inwiefern unterscheiden sich die Studierenden hinsichtlich der Effekte der Präsentationsform auf ihren Lernprozess?

2 Material und Methoden

Für diese Arbeit wurde eine randomisierte, kontrollierte Cross-Over-Studie im Wintersemester 2016/2017 im Rahmen der klinischen Lehre der Universitätsmedizin Göttingen (UMG) durchgeführt, nachdem das Konzept im Sommersemester 2016 pilotiert worden war. Bei der Auswertung wurden qualitative und quantitative Methoden eingesetzt. Der primäre Endpunkt waren Ergebnisunterschiede zwischen Interventions- und Kontrollitems nach Durchführung der Intervention.

Die Grundidee bestand darin, den *Clinical Reasoning*-Lernprozess bei Medizinstudierenden zu fördern, indem sie bei der Bearbeitung bereits etablierter *Key Feature*-Prüfungen mit den häufigsten Falschantworten konfrontiert wurden. Mit diesen Falschantworten sollten die Studierenden sich aktiv auseinandersetzen. Dazu wurde ein *Prompting*-Verfahren – eine Aufforderung zur eigenen Erklärung – mit Freitext-Antwortformat genutzt. Dem aktuellen Kenntnisstand nach sollte dadurch eine tiefere Auseinandersetzung mit dem Inhalt im Sinne von Elaboration erwirkt werden. Um den Lernerfolg zu maximieren, wurden die Studierenden zusätzlich mit einem ausführlichen Feedback unterstützt, welches die richtige Antwort enthielt, sodass diese zeitgleich mit den häufigen Falschantworten zur Diskussion gestellt wurde. Das Ziel war, tiefe Herleitungsprozesse bei der Erklärung und Abgrenzung richtiger und falscher Lösungen zu bewirken, welche im Sinne der weiter oben erwähnten Theorien langfristig bessere Lernleistungen mit sich bringen müsste. Mit einem passenden Studiendesign sollte der Vergleich hergestellt werden zwischen dem Lernerfolg bei beschriebener Reflektion von Falschantworten und traditioneller Beantwortung der *Key Feature*-Fragen.

2.1 Studiendesign

Die hier beschriebene Studie wurde als randomisierte kontrollierte Cross-Over-Studie durchgeführt. Bei diesem Verfahren werden die Probanden randomisiert in zwei Gruppen eingeteilt (Gruppe A und Gruppe B), die beide zu unterschiedlichen Zeitpunkten an der Intervention teilnahmen sowie zu den anderen Zeiten die Kontrollgruppe bildeten.

Der Ablauf der Studie ist in Abbildung 1 dargestellt und wird im Folgenden erläutert.

Den teilnehmenden Studierenden wurden in zehn elektronischen Fallseminaren („E-Fallseminaren“), die als Pflichtveranstaltungen im Stundenplan der Studierenden verzeichnet waren, im wöchentlichen Abstand jeweils vier klinische Fälle im *Key Feature*-Format präsentiert. Diese Fälle beinhalteten je fünf *Key Feature*-Items in Textform. Zu diesen insgesamt 200 Items gehörten 30 Interventionsitems, auf deren Auswertung in dieser Studie besonderes Augenmerk gelegt wurde. Zusätzlich legten die Studierenden einen Eingangs-

test, einen Ausgangstest und einen Retentionstest ab (*Entry Exam*, *Exit Exam* und *Retention Test*), die aus den 30 Interventions-*Key Feature*-Items aufgeteilt auf fünf Fälle bestanden.

Das Cross-Over-Design wurde folgendermaßen angewendet: Da die Intervention in dieser Studie aus 30 *Key Feature*-Items bestand, wurden jeder der beiden Gruppen jeweils 15 Items in Form der Intervention und die anderen 15 Items in Form der Kontrolle dargeboten. Jedes Interventionsitem kam folglich nur in einer Gruppe vor, während die andere Gruppe stattdessen das entsprechende Kontrollitem zu sehen bekam. Zur Verteilung der Interventionsitems auf die beiden Gruppen wurde jedes Item anhand seiner Schwierigkeit gewichtet. Die Schwierigkeit eines Items ergab sich aus seiner Erfolgsquote (Anteil der Studierenden, die das Item richtig beantworteten) in der ersten *Key Feature*-Prüfung (*Entry Exam*), die während der Pilotierung im Sommersemester 2016 stattgefunden hatte. Diese *Key Feature*-Prüfung wurde von allen Studierenden, die sich zu diesem Zeitpunkt im dritten klinischen Semester befanden, absolviert und ihr Ergebnis repräsentiert den Wissensstand der Studierenden zu Beginn des Semesters ohne vorherige Lehre in den abgefragten Themengebieten. Die mittlere Item-Schwierigkeit betrug 0,26. Nach Berechnung der Schwierigkeit jedes Interventionsitems wurden diese so auf die beiden Gruppen verteilt, dass sich für jede Gruppe eine möglichst gleiche gemittelte Item-Schwierigkeit ergab. Nach der Verteilung betrug die mittlere Item-Schwierigkeit für Gruppe A 0,26, für Gruppe B 0,25.

Die im Laufe des Wintersemesters 2016/2017 im wöchentlichen Abstand durchgeführten zehn *Key Feature*-Prüfungen bestanden aus jeweils vier klinischen Fällen mit jeweils fünf Items. Darunter befanden sich die Interventionsitems. Die Gruppen A und B erhielten unterschiedliche Prüfungen, die sich im Vorkommen der Interventionsitems unterschieden. Kam in einer Gruppe ein Interventionsitem vor, befand sich in der anderen Gruppe an dieser Stelle das jeweilige Kontrollitem. Jedes Interventionsitem kam im Laufe der zehn *Key Feature*-Prüfungen zweimal, stets in unterschiedlichen Wochen, jedoch immer innerhalb derselben Gruppe, vor. In einer *Key Feature*-Prüfung gab es pro Gruppe durchschnittlich drei Interventionsitems. Die Studierenden einer Gruppe wurden somit zwischen *Entry Exam* und *Exit Exam* jeweils zweimal mit jedem ihrer 15 Interventionsitems konfrontiert.

Zu Beginn und am Ende des Semesters sowie ein halbes Jahr nach Ende des Semesters wurden die als *Entry Exam*, *Exit Exam* und *Retention Test* bezeichneten *Key Feature*-Prüfungen durchgeführt, in denen jeweils alle 30 Interventionsitems abgebildet waren und die für beide Gruppen gleich waren. Mit diesen Prüfungen sollte der Wissensstand der Studierenden auf dem Gebiet der Interventionsitems zu den drei Zeitpunkten geprüft werden. Die Ergebnisse dieser Prüfungen lieferten unter anderem die Daten für die primäre Analyse (Daten des *Exit Exam*) und für sekundäre Analysen (Daten des *Retention Test*). Der Studienaufbau wird in Abbildung 1 dargestellt.

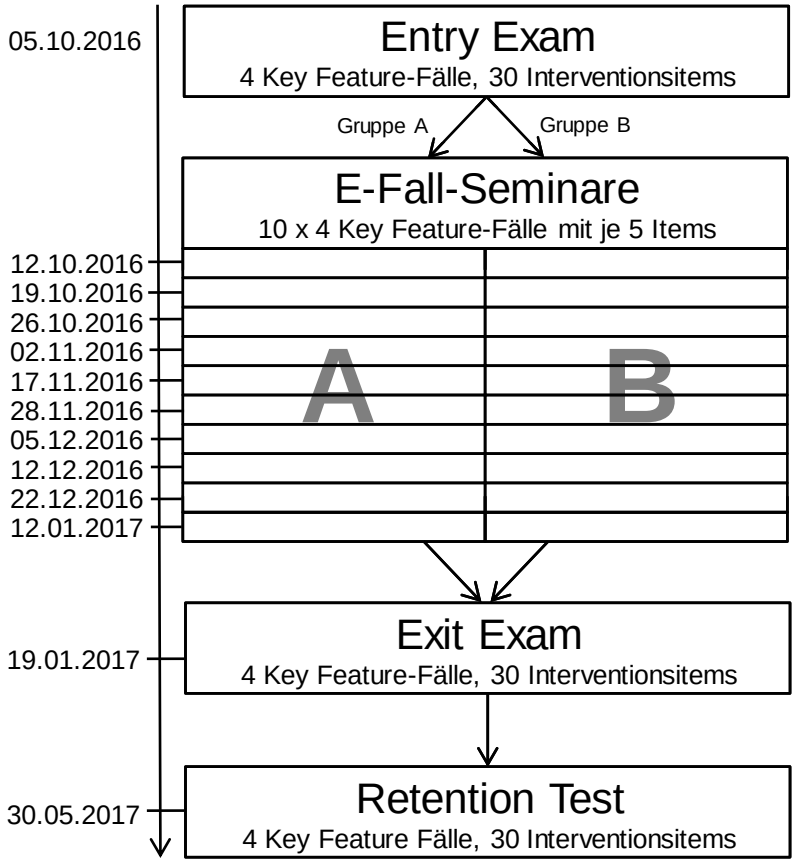


Abbildung 1: Studiendesign und zeitlicher Ablauf

2.1.1 Zeitlicher und inhaltlicher Rahmen

Als Zeitraum für die Intervention wurde das dritte klinische Semester (7. Fachsemester) im Wintersemester 2016/2017 (01.10.2016–31.03.2017) gewählt. Der *Retention Test* fand im Sommersemester 2017 vier Monate nach dem *Exit Exam* statt. Für die genauen Daten siehe Abbildung 1.

Die Lehre im klinischen Studienabschnitt an der Universitätsmedizin Göttingen unterliegt einem modularen Aufbau. Im dritten klinischen Semester werden an der UMG folgende Module nacheinander unterrichtet: Modul 3.1 „Erkrankungen des Herz-Kreislaufsystems und der Lunge“, Modul 3.2 „Erkrankungen der Niere und des Urogenitalsystems“, Modul 3.3 „Erkrankungen des Blutes, des Knochenmarks und Grundlagen der Tumorerkrankungen“ (Stand 07.06.2017). Die E-Fallseminare fanden über alle drei Module hinweg wöchentlich statt.

Die fachlichen Inhalte der zehn *Key Feature*-Prüfungen waren stets an den in der jeweils vorausgegangenen Woche unterrichteten Themen orientiert. So wurde nur Wissen aus den Gebieten abgefragt, die bereits unterrichtet worden waren.

Die Implementierung der *Key Feature*-Prüfungen erfolgte im dritten klinischen Semester, da dort erstmals große Teilbereiche der Inneren Medizin unterrichtet wurden. Die Lerninhalte erschienen für das Lernen und Prüfen von *Clinical Reasoning* besonders geeignet, da die Innere Medizin mit ihrer inhaltlichen Vielfalt und universellen klinischen Relevanz viele wichtige Fallbeispiele für klinische Szenarien liefern kann (Berufsverband deutscher Internisten e.V. 2020). Zudem ergaben sich organisatorische Vorteile dadurch, dass der Leiter des Bereichs Medizindidaktik und Ausbildungsforschung an der UMG gleichzeitig Lehrverantwortlicher für das Modul 3.1 war.

2.1.2 Stichprobe

Initial in die Studie eingeschlossen wurden alle Medizinstudierenden, die sich im Wintersemester 2016/2017 an der Universitätsmedizin Göttingen im dritten klinischen Semester befanden. Diese 153 Studierenden wurden durch Zufallsauswahl, stratifiziert nach Geschlecht und Prüfungsleistung im vorherigen Semester, in zwei gleich große Gruppen eingeteilt. Die beiden Studiengruppen wurden benannt als Gruppe A und Gruppe B.

Studierende, die das dritte klinische Semester nicht vollständig durchliefen, wurden bei der Auswertung aus der Studie ausgeschlossen. Es wurden nur die Daten von Studierenden verwendet, die zu allen für die Studie relevanten Modulen angemeldet waren, die nicht mehr als zwei *Key Feature*-Prüfungen im Laufe des Semesters versäumten und die an *Entry Exam*, *Exit Exam* und *Retention Test* teilnahmen. Außerdem konnten nur die Ergebnisse von Studierenden ausgewertet werden, die alle *Key Feature*-Prüfungen gemäß ihrer Gruppenzuordnung bearbeitet hatten, ansonsten galten die Daten als kontaminiert. Ein weiteres Einschlusskriterium war das Vorliegen der unterschriebenen Einverständniserklärung zur Studienteilnahme (siehe Absatz 2.1.4).

Einschlusskriterien waren die Teilnahme an *Entry Exam*, *Exit Exam* und *Retention Test*, Regelmäßige Teilnahme an den E-Fallseminaren – wobei hier Fehltermine erlaubt waren –, unterschriebene Einverständniserklärung und die fehlerlose Einhaltung der Studiengruppenzuordnung.

2.1.3 Leistungsanreiz

Im Sommersemester 2016 wurde die hier beschriebene Studie als Pretest mit den Studierenden im damaligen dritten klinischen Semester durchgeführt. Das Studiendesign, der inhaltliche und chronologische Aufbau sowie die verwendeten Interventionsitems stimmten mit der späteren Studie überein. Zusätzlich wurden im Sommersemester 2016 nach dem *Exit Exam* Fokusgruppeninterviews durchgeführt. Hier wurde die Meinung der Stu-

dierenden zu den *Key Feature*-Prüfungen erhoben. In den Fokusgruppeninterviews wurden Motivationsdefizite der Studierenden festgestellt. Diese ließen sich auch in den Freitextantworten erkennen. Daraufhin wurde für die Hauptstudie eine Incentivierung durch Büchergutscheine dem Studiendesign hinzugefügt.

Die Studierenden wurden zu Beginn der Studie darüber informiert, dass zur Belohnung für die beste Beantwortung der Freitextfragen Büchergutscheine einer akademischen Buchhandlung vergeben werden würden. Dafür sollte die Summe der Freitextbewertungen über die zehn E-Fallseminare hinweg berücksichtigt werden. Diese Information erhielten die Studierenden, um sie zur ausführlichen Beantwortung der Freitextfragen zu motivieren.

2.1.4 Aufklärung, Ethikvotum, Datenverarbeitung

Bei der Datenerhebung für diese Studie wurden personenbezogene Prüfungsdaten generiert, die in einem späteren Schritt anonymisiert wurden. Dieses Vorgehen wurde vorab durch ein Votum der Ethikkommission der Universitätsmedizin Göttingen genehmigt (Antragsnummer 15/9/16 mit Votum vom 21.09.2016).

Vor Studienbeginn erhielten die Studierenden eine schriftliche Aufklärung über Zweck, Art und Umfang der Datenverarbeitung. Nur bei deren Unterzeichnung wurden die Daten der Studierenden in die Studie eingeschlossen.

Die Prüfungsdaten der Studierenden wurden nach jedem E-Fallseminar aus dem Prüfungsprogramm exportiert und gespeichert. Innerhalb der Prüfungssoftware waren die Daten den Namen und Matrikelnummern der Studierenden zugeordnet. Vor der Auswertung wurden die Daten einer Anonymisierung unterzogen, sodass die Datensätze nicht mehr einzelnen Studierenden zuzuordnen waren (siehe Abschnitt 2.3.1).

2.2 Durchführung

2.2.1 *Key Feature*-Fälle

An der Universitätsmedizin Göttingen werden bereits seit dem Sommersemester 2013 *Key Feature*-Prüfungen im Rahmen elektronischer Fallseminare (E-Fallseminare) für Studierende im klinischen Studienabschnitt durchgeführt. Die dafür entwickelten Fälle kamen von 2013 bis 2015 wiederholt zum Einsatz. Inhalt der Fälle waren internistische Erkrankungen aus den Gebieten Kardiologie, Nephrologie sowie Hämatologie und Onkologie. Ein beispielhafter Fall war folgendermaßen aufgebaut:

Ein einleitender Text (sogenannte „klinische Fallvignette“) beschrieb ein Szenario, in dem sich ein 50-jähriger Raucher, bei dem ein Diabetes mellitus sowie Angina pectoris bekannt sind, mit stärksten thorakalen Schmerzen mit Ausstrahlung in den linken Arm in der Notaufnahme vorstellt. Das erste *Key Feature* war die Frage nach der wahrscheinlichsten Ver-

dachtsdiagnose (akuter Myokardinfarkt). Das zweite *Key Feature* handelte von der Auswahl der richtigen analgetischen Therapie im Zuge der Akutbehandlung. Weitere *Key Feature*-Fragen handelten von der Troponin T-Bestimmung im Labor, dem Einsatz der Koronarangiographie und der Entscheidung für eine perkutane transluminale Koronarangioplastie (PTCA) mit Stentimplantation.

Zwar wurden die Fälle stetig weiterentwickelt, jedoch wurde eine große Kerngruppe an *Key Feature*s und dazugehörigen Items in gleicher Form mehrfach verwendet, sodass die Ergebnisse miteinander verglichen werden konnten. Aus dem Stamm an Fällen, die sich im Laufe dieser Semester bewährt hatten, wurden 40 Fälle für die E-Fallseminare dieser Studie ausgewählt.

Die *Key Feature*-Prüfungen aus den Jahren 2013 bis 2015 fanden ebenfalls jeweils im dritten klinischen Semester unter denselben Rahmenbedingungen statt wie die für diese Studie durchgeführten. Alle Antworten wurden im *Long Menu*-Format gegeben. Die mit den Fragen verknüpften Antwortlisten, auf welche im *Long Menu* zurückgegriffen wurde, waren in allen Semestern die gleichen; es wurden lediglich neue Antworten zu den Listen hinzugefügt. Keine der bisherigen Prüfungen enthielten Freitextfragen.

2.2.2 Entwicklung der Interventionsitems

Die Intervention in dieser Studie war dazu konzipiert, den Lernerfolg bei der Aneignung klinischer Denk- und Handlungsweisen an Schlüsselstellen der Patientenversorgung zu steigern. Dafür mussten zunächst diejenigen Schlüsselstellen ausfindig gemacht werden, für die eine solche Intervention besonders notwendig und erfolgsversprechend erschienen.

2.2.2.1 Auswahl der Interventionsitems

Es wurden dazu die Falschantworten der Studierenden in den E-Fallseminaren der vorherigen Semester analysiert. Um Rückschlüsse über das Antwortverhalten in den bisherigen *Key Feature*-Prüfungen ziehen zu können, wurden die Antworten aller Items aus allen Fällen der bis dahin durchgeführten E-Fallseminare ausgewertet. Es handelte sich dabei um die Ergebnisse aus folgenden Semestern: Sommersemester 2013, Wintersemester 2013/2014, Wintersemester 2014/2015, Sommersemester 2015.

Aus diesen vier Semestern, in denen wöchentliche *Key Feature*-Prüfungen mit den erwähnten Fällen durchgeführt wurden, wurden die Ergebnisse jedes einzelnen Items über die Zeit summiert. Dafür wurden alle richtigen und falschen Antworten für ein Item aus allen vier Semestern in einer gemeinsamen Tabelle dargestellt und die jeweiligen absoluten und relativen Häufigkeiten gleicher Antworten zusammengezählt. Bei der Identifikation besonders fehlerträchtiger Items wurde nicht nur der prozentuale Anteil richtiger Antworten betrachtet, sondern insbesondere auch die Falschantworten in ihrer inhaltlichen Variabilität und relativen Häufigkeit.

Da die meisten Antworten mit einer Reihe von Synonymen vorlagen, wurden zunächst Gruppen aus gleichbedeutenden Antworten gebildet. Die absoluten und relativen Häufigkeiten der einzelnen Antworten wurden innerhalb der Gruppen summiert. Alle als „richtig" klassifizierten Antworten bildeten eine Gruppe, die nicht weiter verändert wurde. Unter den Falschantworten wurden nach der Gruppierung von Synonymen größere Kategorien nach inhaltlicher Bedeutung gebildet. Dabei wurden ähnliche oder unter einem bestimmten Aspekt zusammenpassende Antworten in gemeinsame Kategorien sortiert. Jeder Antwortkategorie wurde ein Name gegeben. So entstanden aussagekräftige Gruppen von Falschantworten im Hinblick auf den jeweiligen Grundgedanken bei der klinischen Entscheidungsfindung. Die Antwortkategorien konnten sich zwischen den einzelnen Items unterscheiden, je nachdem, worauf die Fragestellung eines Items abzielte und welche Bandbreite an Antworten die Frage zuließ.

In Abbildung 2 ist ein Beispiel für die Gruppierung von Falschantworten in Kategorien aufgeführt:

Falschantwort	n
EKG	40
EKG-Kontrolle	4
Kontrolle des EKGs	1
Elektrokardiogramm	3
Elektrogardiographie (EKG) in Ruhe	8
Ruhe-EKG	1
Langzeit-EKG	2
Echokardiographie	10
Herzultraschall	2
Ultraschall des Herzens	2
Sonographie des Herzens	2
Echokardiographie, transthorakal	1
Echokardiographie-TTE	1
Herzkatheter	5
Herzkatheteruntersuchung	4
Koronarangiographie	10

Falschantwort	n
EKG	59
Echokardiographie	18
Koronarangiographie	19

Abbildung 2: Falschantworten zum Item „Ergometrie zur Diagnostik einer relativen Koronarinsuffizienz". Eine 64-jährige Frau stellt sich mit Angina pectoris-artigen Beschwerden in einer Praxis vor. Das Ruhe-EKG ist unauffällig. Die Frage lautete, welche diagnostische Maßnahme ergriffen werden sollte, um dem Verdacht auf eine relative Koronarinsuffizienz nachzugehen. Die richtige Antwort war „Ergometrie". Dargestellt ist ein Ausschnitt aus den Falschantworten mit ihren absoluten Häufigkeiten (n). Synonyme wurden in Absätzen gruppiert, die Häufigkeiten addiert und Antwortkategorien farbig markiert

Mehreren Falschantworten lag ein gemeinsamer Leitgedanke zugrunde. Diese Leitgedanken wurden in der Benennung der größeren Kategorien aufgefasst, wie zum Beispiel das

Grundkonzept „Echokardiographie“ in den Antworten „Herzultraschall“, „Echokardiographie, transthorakal“ und „Echokardiographie, transösophageal“. Durch das Sortieren der Falschantworten in größere Kategorien konnte eine bessere Übersicht darüber erlangt werden, wie Studierende die *Key Feature*-Fragen beantwortet hatten. Zudem ließen sich unter Beachtung der relativen Häufigkeiten Hierarchien der häufigsten Falschantworten zu einem Item erstellen. Dadurch wurde sichtbar, welche Konzepte besonders häufig fehlerhaft verwendet wurden. Wenn besonders viele Studierende eine bestimmte Falschantwort gegeben hatten, wurde daraus geschlossen, dass eine besonders ausgeprägte Fehlannahme unter Studierenden über diese Antwort herrschte, beziehungsweise dass der Falschbeantwortung eine fehlerhafte Unterscheidung zwischen richtiger und falscher Antwort zugrunde lag. Wurde eine Frage von besonders vielen Studierenden falsch beantwortet, jedoch ohne dass eine Falschantwort in ihrer Häufigkeit besonders herausstach, wurde dies als Hinweis auf eine allgemeine Unsicherheit der Studierenden im Umgang mit dem abgefragten Konzept gewertet, oder auf eine nicht gelungene Abgrenzung der richtigen Antwort gegenüber einer Reihe von Falschantworten.

Diese Überlegungen waren die Grundlage für die Auswahl der Items für die Intervention. Nach der Sortierung der Falschantworten in Kategorien wurden 30 Items mit als aussagekräftig empfundenen Falschantworten ausgewählt. Bei diesen ausgewählten Items gab es entweder Antwortkategorien mit besonders hoher relativer Häufigkeit, einen generell geringen Anteil an richtigen Antworten, und/oder das *Key Feature* war von besonders großer klinischer Relevanz (gemessen z.B. an der Häufigkeit des Auftretens im klinischen Alltag oder an der Schwere der Konsequenzen bei fehlerhafter Entscheidung).

2.2.2.2 Gestaltung der Interventionsitems

Die so ausgewählten 30 Items wurden nun für die Intervention modifiziert. Jedes bis dahin in den *Key Feature*-Prüfungen verwendete Item bestand aus einer Beschreibung des Szenarios, einer Fragestellung, welche schlagwortartig im *Long Menu*-Format beantwortet wurde, und einer anschließend gezeigten Feedback-Kombination aus *Knowledge of Correct Response*-Feedback und elaboriertem Feedback (Dempsey 1993) (siehe hierzu Abbildung 4). In diesem Format wurden die Kontroll-Items angezeigt. Bei den Interventionsitems wurden die Beschreibung des Szenarios, die Fragestellung und die Beantwortung im *Long Menu*-Format beibehalten, jedoch wurden die Items jeweils um eine Nachfrage erweitert, die im Freitextformat beantwortet werden sollte, und das kombinierte Feedback wurde aufgeteilt, sodass die beiden Komponenten zeitversetzt gezeigt werden konnten.

Für jedes Interventionsitem wurde eine Freitextfrage formuliert. Dabei wurden die zuvor ermittelten häufigsten und klinisch relevantesten Falschantworten für das jeweilige Item aufgegriffen. Die Freitextfragen enthielten den Auftrag, zu begründen, warum diese Falschantworten nicht korrekt waren.

Der Aufbau eines Interventionsitems sah folgendermaßen aus (siehe hierzu auch Abbildung 5):

Nach Beantwortung der initialen Fragestellung des Items im *Long Menu*-Format wurde zunächst ein reines *Knowledge of Correct Response* (KCR)-Feedback gezeigt, welches die richtige Antwort offenbarte, jedoch keine weiterführende Erklärung enthielt. Somit war für alle Studierenden unabhängig von ihrer Antwort ersichtlich, wie die korrekte Antwort auf die initiale Frage lautete. Daran schloss sich die Nachfrage an.

Die Nachfrage (= Freitextfrage) stellte eine zweite Fragestellung dar, die sich auf die Lösung der ersten Frage bezog und deren Antwort frei formuliert werden musste. Typischerweise bestand die Freitextfrage aus der Aufforderung, die richtige Antwort auf die initiale Fragestellung des Items zu begründen sowie den Unterschied zu einer oder mehreren genannten Falschantworten zu erläutern.

Nach Beantwortung der Freitextfrage wurde ein ausführliches Feedback gezeigt, welches Hintergrundinformationen und Begründungen enthielt. Dieses ausführliche Feedback konnte bei Kontrollitems direkt aufgerufen werden. In der Summe waren also alle Studierenden bei allen Items gegenüber den gleichen Informationen exponiert – mit dem einzigen Unterschied, dass bei Interventionsitems eine zusätzliche Freitextfrage beantwortet werden musste.

2.2.3 Durchführung der *Key Feature*-Prüfungen

Die Intervention und auch die Messung des Lernerfolgs fanden im Rahmen formativer *Key Feature*-Prüfungen statt. Es wurden zwei Formen von *Key Feature*-Prüfungen angewandt, die sich in ihrem Aufbau unterschieden: Einerseits *Entry Exam*, *Exit Exam* und *Retention Test* (zur Messung des Lernerfolgs), andererseits die E-Fallseminare in den Wochen 2 bis 11 (zur Durchführung der Intervention).

Beide Formen von *Key Feature*-Prüfungen wurden auf dieselbe Art und Weise durchgeführt. Die Studierenden befanden sich in Computerräumen, in denen jeder einen eigenen Computer zur Verfügung hatte. Es gab zwei Computerräume von gleicher Größe, auf die die Studierenden je nach Studiengruppe (Gruppe A bzw. Gruppe B) aufgeteilt wurden. Es wurde mit Eingangskontrollen darauf geachtet, dass die Studierenden sich strikt an die Raumaufteilung je nach ihrer Gruppenzuteilung hielten, da die Gruppen A und B unterschiedliche Prüfungen erhielten (ausgenommen *Entry Exam*, *Exit Exam* und *Retention Test*) und die einer Gruppe zugeordnete Prüfung auf alle Computer des für die Gruppe vorgesehenen Computerraumes geladen wurde. Auf jedem Gerät lief die Prüfungssoftware CAMPUS, mittels welcher die *Key Feature*-Prüfungen für die Studierenden dargestellt wurden. Während der Prüfungen war es auf den dafür benutzten Rechnern nicht möglich, das Internet oder andere elektronische Hilfsmittel aufzurufen. Die Studierenden wurden dazu angehalten, die Fragen für sich alleine zu beantworten und nicht in Gruppen zu arbeiten. Zur Bearbeitung einer *Key Feature*-Prüfung standen den Studierenden 45 Minuten zur Ver-

fügung. Hatte ein Studierender nach Ablauf der 45 Minuten noch nicht alle Fragen beantwortet, musste die Prüfung ggf. unvollständig beendet werden.

Bei allen *Key Feature*-Prüfungen handelte es sich um formative Prüfungen. Es wurden keine Leistungspunkte vergeben, die im Zusammenhang mit der Erbringung oder Benotung von Leistungsnachweisen im klinischen Studienabschnitt standen.

2.2.4 Aufbau von *Entry Exam*, *Exit Exam* und *Retention Test*

Die drei Prüfungen vor und nach Durchführung der Intervention (*Entry Exam*, *Exit Exam* und *Retention Test*) waren für beide Gruppen identisch. Es wurden insgesamt 30 Items abgefragt, verteilt auf 5 klinische Fälle mit jeweils 5-7 Items. Alle Interventionsitems kamen in diesen Fällen vor.

Die *Key Feature*-Fälle waren folgendermaßen aufgebaut: Zunächst wurde ein einleitender Text zur Einführung in den Fall dargestellt („Fallvignette"). Dieser Text enthielt Informationen über den zu betrachtenden Patienten, die klinische Problemstellung und den zeitlichen und örtlichen Kontext. Danach wurden dem Teilnehmer 5-7 *Key Feature*s in schriftlicher Form hintereinander präsentiert, die sich in chronologischer Reihenfolge auf den zugrundeliegenden Fall bezogen. Ein Item konnte nur angezeigt und beantwortet werden, wenn das vorherige Item beantwortet und die Eingabe bestätigt wurde. Antworten konnten im Nachhinein nicht mehr geändert werden.

Die *Key Feature*-Items entsprachen folgendem Aufbau (siehe schematische Darstellung in Abbildung 3): In einem kurzen Text wurde das Szenario beschrieben. Daraufhin wurde eine Frage gestellt. Die Antwort darauf wurde im *Long Menu*-Format gegeben. Dabei schrieb der Studierende seine Antwort in ein Textfeld. Nach der Eingabe von mindestens drei Buchstaben wurde unterhalb des Textfeldes ein Menü aufgeklappt, das alle zur eingegebenen Buchstabenkombination passenden hinterlegten Antworten anzeigte. Eine Antwort konnte nun aus dem Menü ausgewählt werden. Die dargestellten Antworten stammten aus Antwortlisten mit mehreren hundert Einträgen, welche im Item Management System (IMS) hinterlegt waren. Nachdem eine Antwort ausgewählt wurde, musste diese Eingabe per Mausklick bestätigt werden. Danach wurde das nächste Item angezeigt. Ein Feedback erschien nicht.

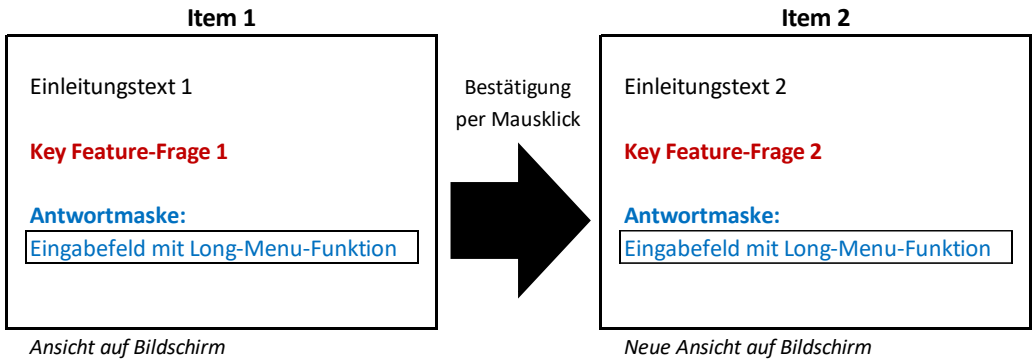


Abbildung 3: Aufbau der *Key Feature*-Items im *Entry Exam*, *Exit Exam* und *Retention Test*. Schematische Darstellung der Anzeige und Chronologie auf den Prüfungsrechnern

2.2.5 Aufbau der E-Fallseminare in den Wochen 2 bis 11

Die Intervention fand im Rahmen von zehn *Key Feature*-Prüfungen statt, die einmal wöchentlich im Rahmen von sogenannten elektronischen Fallseminaren (E-Fallseminaren) durchgeführt wurden. Der Begriff E-Fallseminar wurde als Bezeichnung für dieses Lehrveranstaltungsformat in Abgrenzung zu anderen Lehrveranstaltungen verwendet. Die einzelnen *Key Feature*-Prüfungen enthielten jeweils vier klinische Fälle mit jeweils fünf Items. Die Gruppen A und B erhielten unterschiedliche Prüfungen. Während die klinischen Fälle in beiden Gruppen von Woche zu Woche gleich waren, gab es auf Itemebene Unterschiede hinsichtlich ihrer Darbietung als Interventions- oder Kontrollitem. Von insgesamt 20 Items pro *Key Feature*-Prüfung waren in jeder Gruppe bis zu sechs Interventionsitems, im Durchschnitt waren es drei Interventionsitems pro Gruppe und Woche.

Jeder Fall begann mit einer Fallvignette, die in die klinische Problemstellung und den zeitlichen, örtlichen und persönlichen Kontext einführte. Darauf folgten je fünf *Key Feature*-Items mit entsprechenden Fragestellungen, die analog zu dem Verfahren in *Entry Exam*, *Exit Exam* und *Retention Test* im *Long Menu*-Format beantwortet werden mussten.

Handelte es sich bei einem Item nicht um ein Interventionsitem, wurde im Anschluss an die unwiderrufliche Beantwortung der Frage die Auflösung in Form eines kombinierten Feedbacks gezeigt. Dieses bestand einerseits aus einem *Knowledge of Correct Response* (KCR)-Feedback: einem Satz, welcher kurz und knapp die Antwort auf die vorherige Frage gab und an erster Stelle auf der nachfolgend angezeigten Seite stand. Andererseits bestand es aus einem elaborierten Feedback in Form eines längeren Texts, welcher in einem Textkasten an zweiter Stelle auf der neuen Seite stand und durch Anklicken vergrößert werden konnte. Nur in vergrößerter Form war der Text im Kasten ausreichend lesbar. Somit war es den Probanden freigestellt, das Feedback durch Anklicken zu vergrößern und durchzulesen. Aus technischen Gründen war es nicht möglich, zu erfassen, welches Feedback von

welchen Studierenden gelesen wurde und wie lang der Text jeweils angezeigt wurde. Der elaborierte Feedback-Text enthielt eine Begründung der richtigen Antwort, zudem wurden mögliche Falschantworten genannt und von der richtigen Antwort abgegrenzt. Zuletzt enthielt der Textkasten eine Auflistung aller Antworten, die in der jeweiligen Frage als richtig gewertet wurden. Im Anschluss an das Feedback waren das Szenario und die Fragestellung des nächsten Items zu sehen. Siehe Abbildung 4 zur Verdeutlichung des Aufbaus.

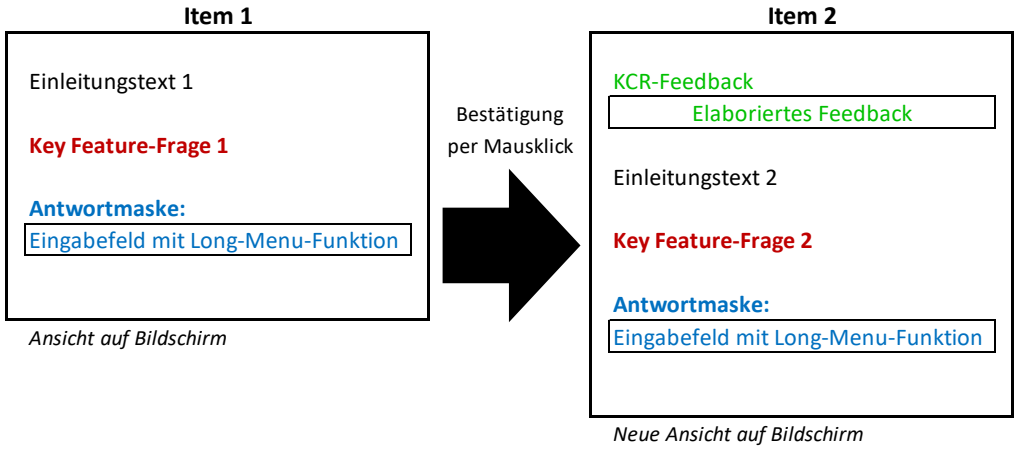


Abbildung 4: Schematische Darstellung des Aufbaus und der Chronologie eines Kontrollitems in den *Key Feature*-Prüfungen der Wochen 2-11

Die Kontrollitems unterschieden sich in ihrem Aufbau nicht von anderen Nicht-Interventionsitems.

Die Interventionsitems hatten einen abweichenden Aufbau. Dieser wurde bereits im Abschnitt 2.2.2.2 beschrieben und wird hier kurz im Vergleich zu den Kontrollitems dargestellt. Zunächst wurde derselbe kurze Text dargeboten und dieselbe Frage wie im korrespondierenden Kontrollitem gestellt, welche ebenfalls im *Long Menu*-Format beantwortet werden musste. Anstelle des kompletten Feedbacks wurde im Anschluss jedoch nur das KCR-Feedback angezeigt als kurzer Satz und mit Auflistung der als richtig gewerteten Antworten im Textkasten zum Anklicken und Vergrößern. Daraufhin wurde den Studierenden die Nachfrage präsentiert und ein Freitextfeld zur Eingabe der Antwort. Anders als in den *Long Menu*-Items war hier keine Antwortliste mit dem Eingabefeld verknüpft, sondern es konnten beliebige Zeichen eingegeben werden. Nach Eingabe und Bestätigung ihrer Freitextantwort bekamen die Studierenden die nächste Ansicht mit dem elaborierten Feedback im Textkasten angezeigt. Der Feedbacktext entsprach inhaltlich dem in den Kontrollitems. Auf das elaborierte Feedback folgte wie in den anderen Items das Szenario des nächsten Items. In der folgenden Abbildung wird der Aufbau der Interventionsitems schematisch dargestellt.

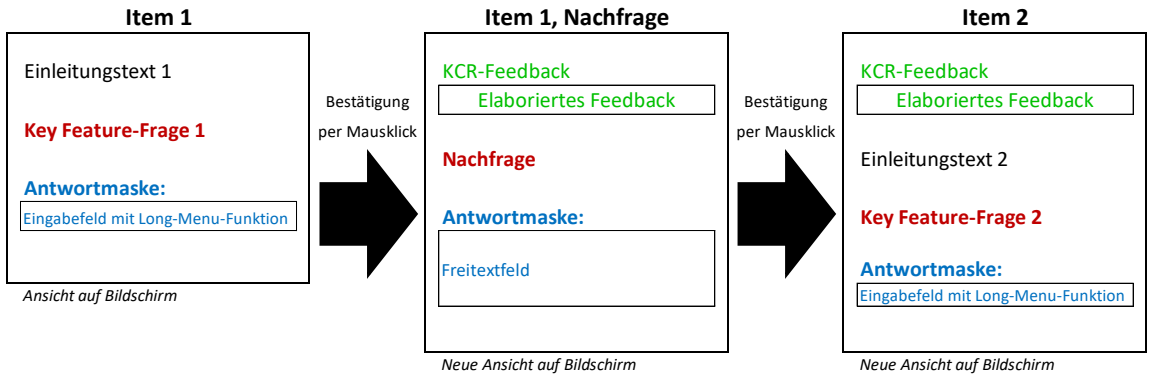


Abbildung 5: Schematische Darstellung des Aufbaus und der Chronologie der Interventionsitems

2.3 Auswertung

2.3.1 Datenausgabe und -aufbereitung

Die bei jedem Termin gespeicherten Daten jedes teilnehmenden Studierenden beinhalteten den Anmeldenamen im universitätseigenen IT-System StudIP („StudIP-Anmeldename“), den Vor- und Nachnamen und die Matrikelnummer. Außerdem wurde die Zeit erfasst und gespeichert, die der Studierende zum Bearbeiten der Fragen brauchte (Beginn der Zeitmessung am Prüfungsstart, Ende der Zeitmessung bei eigenhändiger Bestätigung der Abgabe oder Beendigung der Prüfung durch die Aufsichtskraft). Es wurden zudem alle eingegebenen Antworten zu den einzelnen Items gespeichert.

Die Prüfungsdaten aus allen *Key Feature*-Püfungen wurden zu longitudinalen Datensätzen zusammengefasst und diese in einer gemeinsamen Tabelle dargestellt. Die Prüfungsdaten der Studierenden wurden dann vom Studienleiter anonymisiert, indem Name, StudIP-Anmeldename und Matrikelnummer aus der Datei gelöscht wurden. Mit diesem Datensatz wurde in Microsoft Excel und im Statistikprogramm SPSS weitergearbeitet.

2.3.1.1 *Entry Exam*, *Exit Exam* und *Retention Test*

Die Aufbereitung der Daten aus *Entry Exam*, *Exit Exam* und *Retention Test* unterschied sich von der der Daten aus den dazwischen liegenden E-Fallseminaren. Die Datensätze der erstgenannten drei Prüfungen wurden in zwei Formen für die nachfolgende Analyse vorbereitet:

Zunächst wurden die Ergebnisse aller Items dichotomisiert dargestellt. Dabei konnte jedes Item zwei Werte annehmen: Richtig (1) oder falsch (0). Alle inhaltlich korrekten Antworten

auf ein Item wurden als „richtig“ klassifiziert und alle Falschantworten als „falsch“. Die Klassifikation erfolgte zweistufig: Da es sich um *Long Menu*-Items handelte, gab es eine im Prüfungssystem hinterlegte Liste richtiger Antworten, mit der eine automatische Auswertung erfolgte. Alle nicht zu dieser Liste gehörigen Antworten wurden als falsch klassifiziert. Zusätzlich wurden alle Antworten von der Autorin manuell überprüft und gegebenenfalls Antworten im Nachhinein umklassifiziert, die bei der automatischen Auswertung falsch eingeordnet wurden, beispielsweise weil eine korrekte Antwort nicht in der vordefinierten Liste eingeschlossen war.

In einer zweiten Ergebnisaufbereitung wurden die Falschantworten in mehrere Kategorien unterteilt. Alle korrekten Antworten erhielten weiterhin den Wert „richtig“ („1“). Für die Falschantworten gab es bis zu sechs spezifische Kategorien („2“ bis „7“) sowie die unspezifische Kategorie „0“, in die alle nicht klassifizierten Falschantworten sortiert wurden. Die Einteilung wurde, ähnlich wie die in Abschnitt 2.2.1 dargelegte, nach Bedeutung und inhaltlicher Zusammengehörigkeit der Antworten vorgenommen. Die Kategorisierung der Falschantworten war für jedes Item unterschiedlich, aber für die korrespondierenden Items in *Entry Exam*, *Exit Exam* und *Retention Test* untereinander identisch. So wurde zwischen den drei Prüfungen eine Vergleichbarkeit auf Ebene der Falschantworten geschaffen.

2.3.1.2 E-Fallseminare 2-11

Die Datensätze der E-Fallseminare in den Wochen 2-11 wurden folgendermaßen aufbereitet:

Die Ergebnisse aller *Long Menu*-Items wurden dichotomisiert als richtig (1) oder falsch (0) dargestellt. Für jedes *Long Menu*-Item gab es eine vordefinierte Liste richtiger Antworten, die zur automatischen Auswertung herangezogen wurde. Alle nicht zu dieser Liste gehörigen Antworten wurden als falsch klassifiziert. Zusätzlich wurden alle Antworten von der Autorin manuell überprüft und gegebenenfalls Antworten im Nachhinein umklassifiziert, die bei der automatischen Auswertung falsch eingeordnet wurden, etwa weil eine korrekte Antwort nicht in der vordefinierten Liste eingeschlossen war.

Zudem wurden die Freitexte, welche bei den Interventionsitems eingegeben worden waren, einer inhaltlichen Analyse unterzogen und auf einer dreistufigen Skala bewertet. Die Skala umfasste die Werte 2, 1 und 0. Eine Freitextantwort wurde mit „0“ bewertet, wenn sie keine Antwort auf die Frage enthielt, inhaltlich das Thema verfehlte oder grob falsch war. Mit „1“ wurden Freitextantworten bewertet, die sich mit den in der Frage angesprochenen Mechanismen auseinandersetzten, erkennbar fachlich passende Gedanken repräsentierten, jedoch keine vollständige, ausreichend präzise und korrekte Antwort darstellten. Mit „2“ wurden schließlich alle korrekten Antworten bewertet.

Diese allgemeingültige Definition der Werte 2, 1 und 0 wurde durch konkret spezifizierte Definitionen für jedes Interventionsitem ergänzt. Dabei wurde die Individualität der Interventionsitems in ihren Fragestellungen berücksichtigt. Für jedes Item wurde ein Erwartungshorizont für die Bewertung mit 2 oder 1 festgelegt, in dem Kernaussagen, Schlagworte oder Themen definiert wurden, die mindestens in einer Antwort vorkommen mussten, damit diese mit 2 beziehungsweise 1 bewertet werden konnte. Dazu wurden Ankerbeispiele hinterlegt. Die Voraussetzungen für eine 0-Punkte-Antwort blieben für jedes Item dieselben. Tabelle 1 zeigt die itemspezifische Definition der Bewertungskategorien mit dazu passenden Ankerbeispielen anhand eines beispielhaften Items.

Tabelle 1: Kriterien für die Bewertung von Freitextantworten am Beispiel des Interventionsitems "Erkennen einer Tachyarrhythmia absoluta im EKG". In den mit „Beispiel“ beschrifteten Spalten befinden sich unveränderte Zitate aus den Freitextantworten der Studierenden

Item: Erkennen einer Tachyarrhythmia absoluta im EKG (E-Fallseminar 4, Fall 08, Key Feature 1)					
Nachfrage: Bitte erläutern Sie kurz, woran erkennbar ist, dass in diesem EKG am ehesten eine Tachyarrhythmia absoluta zu sehen ist.					
2 Punkte		**1 Punkt**		**0 Punkte**	
Definition	Beispiel	Definition	Beispiel	Definition	Beispiel
Fehlen von p-Wellen ODER Unregelmäßigkeit der QRS-Komplexe (mind. eines von beidem)	„Unregelmäßige QRS-Komplexe mit einer Frequenz über 100/min, keine P-Wellen erkennbar.“	Fehlerhafte oder unvollständige Analyse des EKGs bzw. fehlerhafte Schlussfolgerung. Verwendung von Fachbegriffen	„Kein regelrechter SR und HF über 100“	Keine Angabe/Thema verfehlt/komplett falsche Antwort	„weiß nicht“

Bei der Bewertung der Freitexte wurde sowohl auf die Nennung von Schlüsselbegriffen als auch auf deren sinnvolle Einbettung in den Text geachtet. Die unterschiedlichen Anforderungen, die an die Freitextantworten gestellt wurden, richteten sich stets nach Art und Formulierung der Nachfrage. So gab es Fragen, in denen lediglich die Aufzählung von Fakten gefordert wurde (Beispiele: „Bitte zählen Sie die 4 Kriterien auf, die in den CRB-65-Score eingehen.“ (E-Fallseminar 6, Fall 11, *Key Feature* 4); „Wie unterscheidet sich ein Nephrotisches Syndrom von einem Nephritischen Syndrom?“ (E-Fallseminar 7, Fall N04, *Key Feature* 1)), aber auch Fragen, die eine Erläuterung von Prinzipien erforderten (Beispiele: „Warum wäre eine Röntgenuntersuchung des Thorax in dieser Situation eher nicht zielführend?“ (E-Fallseminar 6, Fall 12, *Key Feature* 3); „Bitte begründen Sie, warum bei dem Ver-

dacht auf eine restriktive Lungenerkrankung die Bestimmung der Vitalkapazität zur Diagnosestellung allein nicht ausreicht." (E-Fallseminar 8, Fall 15, *Key Feature* 3)). Bei letzteren Fragen reichte die alleinige Nennung eines Schlagwortes nicht aus, um eine korrekte Antwort zu bilden, sondern die Antwort musste in Form eines überzeugenden Argumentes dargelegt werden. Nachfragen, die zwei Beantwortungsschritte forderten (Beispiel: „Bitte erläutern Sie kurz, was ein 2-Kammer-Schrittmacher ist und wodurch er sich von einem biventrikulären Schrittmacher unterscheidet" (E-Fallseminar 4, Fall 14, *Key Feature* 5)), wurden nur mit zwei Punkten bewertet, wenn beide Sachverhalte in der Freitextantwort korrekt erklärt wurden.

Die Bewertung der Freitextantworten wurde von der Autorin vorgenommen, zusätzlich wurde stichprobenartig eine Validitätsprüfung durch einen Facharzt für Innere Medizin vorgenommen. Die vorherige Erarbeitung der Bewertungsgrundlage erfolgte ebenfalls durch die Autorin unter sorgfältiger fachärztlicher Supervision.

2.3.1.3 Modulklausurdaten

Zusätzlich zu den Ergebnissen der in der Studie untersuchten *Key Feature*-Prüfungen wurden die Ergebnisse der curricularen Modulklausuren herangezogen. Es handelte sich um die summativen Prüfungen der Module des vorangegangenen Semesters (2. klinisches Semester im Sommersemester 2016) und des Semesters, in dem die Studie stattfand (3. klinisches Semester im Wintersemester 2016/2017). Die Klausurdaten wurden vom Studiendekanat der medizinischen Fakultät bereitgestellt. Es wurden Prozentscores für jeden Studierenden gebildet (Anteil der erreichten Klausurpunkte an den maximal erreichbaren Punkten) und die Scores der einzelnen Klausuren zusammengefasst zu Gesamtscores über ein gesamtes Semester (Sommersemester 2016) bzw. über alle Klausurfragen der Inneren Medizin (Wintersemester 2016/2017).

2.3.2 Datenanalyse

Wie die entstandenen Daten zur Beantwortung der Forschungsfragen und zur Beschreibung von Stichprobe und Material ausgewertet wurden, wird im folgenden Abschnitt dargestellt. Alle statistischen Analysen wurden mit dem Computerprogramm SPSS 14.0 (IBM) durchgeführt, unterstützend kam Microsoft Excel 2010 (Microsoft Office 2010) zum Einsatz. Das Signifikanzniveau wurde auf 5% ($p < 0{,}05$) festgelegt. Sofern nicht anders angegeben, wurden die erhobenen Daten als Mittelwert ± Standardabweichung dargestellt.

2.3.2.1 Deskriptive Statistik

Zur Beschreibung der Stichprobe wurden demographische Daten der teilnehmenden Studierenden und Prozessdaten, beispielsweise die Bearbeitungszeiten der E-Fallseminare, hinsichtlich Spannweite, Mittelwert und Standardabweichung ausgewertet. Für *Entry Exam*, *Exit Exam* und *Retention Test* erfolgte eine Itemanalyse, bei der die Itemschwierigkeit und –

trennschärfe jedes Items bestimmt wurde. Außerdem wurde jeweils die interne Konsistenz mithilfe des Cronbachschen Alpha (Cortina 1993) berechnet.

2.3.2.2 Forschungsfrage 1: Vergleich des Lernerfolgs in den Interventions- und Kontrollitems

Aus den dichotomisierten Antworten des *Entry Exam*, *Exit Exam* und *Retention Test* wurden zunächst Summen gebildet. Dabei wurde die Gesamtpunktzahl (entspricht der Anzahl der richtigen Antworten) eines jeden Studierenden in jeder der drei Prüfungen errechnet (*Score*). Aus dieser Summe wurde der Anteil der erreichten Punktzahl an der maximal erreichbaren Punktzahl in Prozent errechnet („Prozentscore"). Die Prüfungsergebnisse wurden für die statistische Analyse zudem einem *Dummy Coding*-Verfahren unterzogen. Dafür wurden analog zum eben beschriebenen Vorgehen die Summen der von den Studierenden erreichten Punkte in ihren jeweiligen Interventions- und Kontrollitems getrennt berechnet, passend zu ihrer jeweiligen Gruppenzugehörigkeit. In jedem longitudinalen Datensatz konnte so das Ergebnis in den Interventions- und Kontrollitems getrennt voneinander betrachtet werden. Auch aus diesen Summen wurden Prozentscores gebildet, die den Anteil der erreichten Punkte an den maximal erreichbaren Punkten darstellten. Diese Zahlen wurden in eigenen Variablen für jede der drei Prüfungen präsentiert. In einer neuen Variable wurde also der prozentuale Anteil der erreichten Punkte in den für den jeweiligen Studierenden nach seiner Gruppeneinteilung geltenden Interventionsitems oder Kontrollitems im *Entry Exam*, *Exit Exam* oder *Retention Test* gezeigt.

Mithilfe des *Dummy Codings* konnten die Ergebnisse in den Interventionsitems im Vergleich zu den Kontrollitems ausgewertet werden.

Die erste Forschungsfrage bezog sich auf die Veränderung der *Clinical Reasoning*-Kompetenz durch die Intervention. Der primäre Endpunkt der Studie war somit der Lernerfolg in den Interventionsitems am Ende des Semesters, gemessen am Anteil richtiger Antworten im *Exit Exam*. Um herauszufinden, ob der Lernerfolg in den Interventionsitems größer war als in den Kontrollitems, wurden die Ergebnisse in den beiden Kategorien mit einem gepaarten T-Test verglichen. Dafür wurden die mittels *Dummy Coding* vorbereiteten prozentualen Anteile richtig beantworteter Interventionsitems und Kontrollitems im *Exit Exam* verwendet.

Als sekundärer Endpunkt wurden auch die korrespondierenden Ergebnisse des *Retention Tests* betrachtet. Auch hier wurden die Prozentscores in den Interventionsitems und in den Kontrollitems mit einem gepaarten T-Test verglichen.

2.3.2.3 Forschungsfrage 2: Ergebnisunterschiede zwischen den Studierenden anhand unterschiedlicher Charakteristika

Um Aussagen darüber zu treffen, inwiefern die Studierenden unterschiedlichen Alters, Geschlechts und Leistungsstands sich hinsichtlich des Lernerfolgs durch die Intervention un-

terschieden, wurde zunächst der individuelle Lernerfolg durch die Intervention als Differenz zwischen den Prozentscores in den Interventionsitems und in den Kontrollitems berechnet. Grundlage waren die jeweiligen Prozentscores im *Exit Exam* und im *Retention Test*. Danach wurde mittels linearer Regression der Zusammenhang zwischen dem individuellen Lernerfolg als abhängiger Variable und Geschlecht, Alter oder Leistungsstand als unabhängige Variablen berechnet. Als Variablen für den Leistungsstand wurden die Ergebnisse (Prozentscores) in den Modulklausuren des vorausgegangenen Semesters (Sommersemester 2016) sowie die Prozentscores in den Modulklausurfragen der Inneren Medizin im Wintersemester 2016/2017 genutzt.

Es wurden ebenfalls Umfang und Qualität der Freitextantworten der Studierenden ausgewertet. Die Länge der einzelnen Freitextantworten wurde in Zeichen gezählt. Auf Studierendenebene wurde eine Gesamtsumme der Zeichenanzahl in allen Freitextantworten zusammen berechnet („Freitextlänge"). Anhand der qualitativen Bewertungen der Freitexte (siehe Abschnitt 2.3.1.2) wurden Punktsummen berechnet („Freitextscores").

Für jede unabhängige Variable (die genannten Variablen für den Leistungsstand sowie für Umfang und Qualität der Freitextantworten) wurde eine bivariate lineare Regression mit dem individuellen Lernerfolg durch die Intervention als abhängiger Variable durchgeführt. Zudem wurde eine multiple Regression unter Einschluss aller unabhängigen Variablen und derselben abhängigen Variable durchgeführt. Es wurde davon ausgegangen, dass einige der unabhängigen Variablen miteinander korrelieren – beispielsweise die Variablen für den Leistungsstand (die Prozentscores in den Modulklausuren des vorangegangenen Semesters sowie in den Klausurfragen für Innere Medizin im Studiensemester), aber auch die Variablen für Qualität und Länge der Freitextantworten. Korrelierten diese Variablen miteinander, wären sie nicht linear unabhängig voneinander und nähmen sich daher im Regressionsmodell gegenseitig an Effektstärke. Deshalb wurde in weiteren multiplen Regressionen jeweils eine dieser Variablen fallengelassen, um die volle Effektstärke der anderen Variable besser messen zu können.

Als Maß für die Effektstärke wurden der Koeffizient B und der standardisierte Koeffizient Beta herangezogen, sowie als Maß für die Signifikanz der p-Wert.

2.3.2.4 Explorative Analysen

Longitudinale Analyse des Antwortverhaltens

Um den Lernerfolg in den Interventionsitems in Abhängigkeit von der qualitativen Bewertung der Freitextantworten zu analysieren, wurde der longitudinale Verlauf des Antwortverhaltens jedes Probanden für jedes seiner gruppenspezifischen Interventionsitems ermittelt. Dabei wurden die erstmalige Präsentation eines Interventionsitems in einem E-Fallseminar, die sich daran anschließende Freitextfrage, die zweite Präsentation des Interventionsitems in einem späteren E-Fallseminar, die zweite Freitextfrage und die Abfrage des Items im *Exit Exam* berücksichtigt. Die Ergebnisse des *Entry Exams* und des *Retention*

Tests fehlen aus Gründen der Übersichtlichkeit, da die Ergebnisse in einem Baumdiagramm dargestellt wurden. Die Bewertung der Freitextantworten wurde für diese Analyse dichotomisiert. Dabei wurden alle auf der dreistufigen Skala mit „0“ bewerteten Freitextantworten (siehe oben) weiterhin als „0“ (grob falsch oder keine ernsthafte Antwort) klassifiziert, während die mit „1“ und mit „2“ bewerteten Freitextantworten gemeinsam als „1“ (ernsthafte Antwort) klassifiziert wurden.

So entstand für jeden Studierenden pro Interventionsitem ein Verlauf mit fünf dichotomen Verzweigungen. Diese Verläufe wurden zusammengezählt, sodass ein gemeinsamer, Probanden- und Item-unspezifischer Verlauf entstand. Dieser stellte auf allen Stufen der fünf dichotomen Verzweigungen die Gesamtzahl der Studierenden dar, die in einem Interventionsitem die jeweilige Antwortqualität abgegeben hatte. Aus diesen absoluten Zahlen wurden relative Anteile in Prozent errechnet, um darzustellen, welcher Anteil der Studierenden im Durchschnitt ein jeweiliges Antwortverhalten zeigte. Die beschriebenen Prozentzahlen wurden in Form eines fünffach dichotom verzweigten Diagrammes graphisch dargestellt.

Häufigkeiten von Falschantworten

Um zu erfassen, wie sich das Antwortverhalten hinsichtlich des Musters an Falschantworten veränderte, wurden alle Antworten, die zu den verschiedenen Messzeitpunkten auf die 30 Interventionsitems gegeben wurden, für jedes Interventionsitem und jeden Zeitpunkt einzeln gezählt und kategorisiert (wie bereits in Abschnitt 2.3.1.1 beschrieben). Die einzelnen Antworten mit ihren relativen Häufigkeiten wurden innerhalb ihrer Kategorien summiert. Alle nicht kategorisierten Antworten wurden als „andere/keine Antwort“ klassifiziert und ihre Häufigkeiten summiert.

3 Ergebnisse

In diesem Kapitel werden die im Rahmen dieser Arbeit erhobenen Daten dargestellt, die zur Beantwortung der Forschungsfragen relevant sind. Außerdem werden für die weitere Interpretation relevante Daten gezeigt und es wird ein deskriptiver Überblick über die Stichprobe und die verwendeten Items gegeben.

3.1 Charakterisierung der Stichprobe

Die Kohorte der Studierenden, die im Wintersemester 2016/2017 das dritte klinische Semester durchliefen und damit für die Studie infrage kamen, umfasste 153 Studierende. Davon wurden 108 Studierende in die Studie eingeschlossen, die Datensätze der restlichen 45 Studierenden wurden nicht ausgewertet.

18 Studierende hatten keine Einverständniserklärung für die Teilnahme an der Studie gegeben, 2 Studierende mussten wegen der Kontamination ihrer Daten durch nicht plangemäßes Wechseln der Studiengruppe von der Auswertung ausgeschlossen werden. Von den restlichen nicht in die Studie eingeschlossenen Studierenden lagen durch Nichtteilnahme am Eingangs-, Ausgangs- oder Retentionstest keine vollständigen longitudinalen Datensätze vor. Die Rücklaufquote betrug damit 70,6%.

In der 108 Studierende umfassenden Stichprobe betrug das Durchschnittsalter am ersten Tag des Wintersemesters 2016/2017 24,4 ± 2,8 Jahre. Der Anteil weiblicher Studierender betrug 65,7% (71 Studierende). Der durchschnittliche Erfolg in den Modulklausuren des vorangegangenen Semesters betrug 79,8% ± 7,9 (Anteil der zu erreichenden Punkte; gemittelt über alle Klausuren im genannten Zeitraum und alle Studierende in der Stichprobe), in den Modulklausuren des Wintersemesters 2016/2017 80,2% ± 6,7, in den Klausurfragen aus der Inneren Medizin im Wintersemester 2016/2017 76,0% ± 9,1.

3.1.1 Teilnahme an den E-Fallseminaren

Von den 108 in die Studie eingeschlossenen Studierenden nahmen gemäß Einschlusskriterien jeweils 100% an *Entry Exam*, *Exit Exam* und *Retention Test* teil. Die mittlere Anwesenheit in den E-Fallseminaren der Studienwochen 2 bis 11 betrug 91,2%. Die Anwesenheit bei den einzelnen Terminen reichte von 78,7% (in Studienwoche 10) bis 100% (in Studienwoche 6). In Tabelle 2 sind die einzelnen Werte dargestellt.

Tabelle 2: Anwesenheit bei den E-Fallseminaren

Anwesenheit bei den E-Fallseminaren	
E-Fallseminar	Anwesende Studierende (n = 108)

Entry Exam	100,0%
Woche 2	87,0%
Woche 3	98,1%
Woche 4	96,3%
Woche 5	90,7%
Woche 6	100,0%
Woche 7	94,4%
Woche 8	87,0%
Woche 9	84,3%
Woche 10	78,7%
Woche 11	95,4%
Exit Exam	100,0%
Retention Test	100,0%
Durchschnittliche Anwesenheit Wochen 2-11	92,7%
Durchschnittliche Anwesenheit Gesamtzeitraum	93,2%

3.1.2 Eigenschaften der Studiengruppen A und B

Von den 108 in die Studie eingeschlossenen Studierenden gehörten 53 (49,1%) der Studiengruppe A an, 55 (50,9%) der Studiengruppe B. Das Durchschnittsalter der Studierenden betrug in Gruppe A 24,5 ± 2,9 Jahre, in Gruppe B 24,4 ± 2,6 Jahre ($p = 0{,}786$). Der Anteil weiblicher Studierender lag in Gruppe A bei 67,9%, in Gruppe B bei 63,6% ($p = 0{,}639$). Hinsichtlich ihrer Klausurergebnisse im vorausgegangenen Semester (Sommersemester 2016) gab es zwischen den Gruppen A und B keinen signifikanten Unterschied.

3.2 Charakterisierung des Messinstruments: Itemkennwerte

Die anhand der Prüfungsergebnisse ermittelten Werte zur Itemschwierigkeit und Itemtrennschärfe sind in Anhang 1 bis Anhang 3 dargestellt. Die Itemschwierigkeit reichte im *Entry Exam* von 0,01 bis 0,63. Im *Exit Exam* variierte sie zwischen 0,19 und 0,85 und im *Retention Test* zwischen 0,09 und 0,85. Die Itemtrennschärfe reichte von -0,246 bis 0,574, wobei sie im *Entry Exam* im Mittel geringer ausfiel (arithmetisches Mittel: 0,184) als im *Exit Exam* (0,381) und *Retention Test* (0,383). Auch für die Itemschwierigkeit ergaben sich im *Entry Exam* geringere Werte (arithmetisches Mittel: 0,273) als im *Exit Exam* (0,640) und im *Retention Test* (0,602).

Die interne Konsistenz der *Key Feature*-Prüfungen wurde anhand des Cronbachschen Alpha gemessen. Tabelle 3 zeigt die einzelnen Werte.

Tabelle 3: Interne Konsistenz der *Key Feature*-Prüfungen

	Cronbachsches Alpha
Entry Exam	0,622
Exit Exam	0,857
Retention Test	0,859

3.3 Forschungsfrage 1: Effekte der Intervention auf den Lernerfolg

Forschungsfrage 1 zielte darauf ab, welche Auswirkungen die Intervention auf den *Clinical Reasoning*-Lernerfolg hatte. Zur Beantwortung dieser Frage werden die folgenden Ergebnisse betrachtet, die auch in

Tabelle 4 dargestellt werden.

Der mittlere Anteil richtig beantworteter Items betrug für die gesamte Stichprobe (n = 108) und für alle 30 Items im *Entry Exam* 27,3 ± 11,7%. Im *Exit Exam* wurden im Mittel 64,0 ± 19,9% der Items richtig beantwortet, im *Retention Test* waren es 60,2 ± 20,1%. Über alle Items hinweg zeigte sich somit ein allgemeiner Lernerfolg, der zum Zeitpunkt des *Exit Exam* etwas größer war als zum Zeitpunkt des *Retention Test*.

Bei der Betrachtung der Ergebnisse des *Exit Exam* zeigte sich ein statistisch signifikanter Unterschied zwischen Interventions- und Kontrollitems. Anhand der Daten in

Tabelle 4 ist zu sehen, dass die Interventionsitems im *Exit Exam* signifikant häufiger richtig beantwortet wurden als die Kontrollitems. Die Effektstärke für diesen Unterschied betrug d = 0,16. Im *Entry Exam* zeigten die korrespondierenden Daten keinen signifikanten Unterschied. Neben einem allgemeinen Lernerfolg in allen Items ist somit am Ende des Interventionszeitraumes in den Interventionsitems ein größerer Lernerfolg als in den Kontrollitems zu verzeichnen. In Gruppe A zeigte sich dieser Unterschied am deutlichsten. Dort wurden die Interventionsitems im *Exit Exam* zu 65,8 ± 20,8% richtig beantwortet und die Kontrollitems zu 59,8 ± 24,9% (p = 0,008). In Gruppe B betrug der Anteil richtiger Antworten in den Interventionsitems 65,7 ± 18,7%, in den Kontrollitems 64,9 ± 20,6% (p = 0,659).

Im *Retention Test* zeigte sich dieser Unterschied zwischen Interventions- und Kontrollitems nicht mehr.

Tabelle 4: Ergebnisunterschiede zwischen Interventions- und Kontrollitems

	Anteil richtiger Antworten in Prozent			
	Alle Items	Interventionsitems	Kontrollitems	p
Entry Exam	27,3 ± 11,7	27,4 ± 13,1	27,1 ± 14,7	0,832
Exit Exam	64,0 ± 19,9	65,7 ± 19,6	62,4 ± 22,9	**0,022**
Retention Test	60,2 ± 20,1	60,1 ± 21,9	60,3 ± 21,2	0,934

3.4 Forschungsfrage 2: Ergebnisunterschiede zwischen den Studierenden unterschiedlicher Charakteristika

Forschungsfrage 2 zielte darauf ab, zu untersuchen, inwiefern die Studierenden unterschiedlichen Alters, Geschlechts und Leistungsstands sich hinsichtlich ihres individuellen Lernerfolgs durch die Intervention unterschieden. Die Ergebnisse der bivariaten und multiplen linearen Regressionsanalysen, bei denen jeweils die Prozentscore-Differenz zwischen Interventions- und Kontrollitems als abhängige Variable und verschiedene Charakteristika der Studierenden als unabhängige Variablen untersucht wurden, werden im Folgenden präsentiert.

Die Tabellen 5 und 6 zeigen die Ergebnisse der multiplen Regression unter Einbeziehung aller unabhängiger Variablen. Tabelle 7 zeigt die Ergebnisse der multiplen Regression für die Ergebnisse des *Retention Tests* unter Auslassung einer unabhängigen Variable, der erreichten Punktzahl in den Klausuren des Vorsemesters.

3.4.1 Unterschiede nach Geschlecht

In der bivariaten Regression hatten Studierende weiblichen Geschlechts eine um durchschnittlich 2,15 Prozentpunkte größere Prozentscore-Differenz zwischen Interventions- und Kontrollitems im *Exit Exam* als männliche Studierende (Koeffizient B = 2,15; Beta = 0,07; p = 0,488). Im *Retention Test* war die Differenz bei weiblichen Studierenden um durchschnittlich 7,21 Prozentpunkte größer und signifikant (B = 7,21; Beta = 0,22; p = 0,020). Für das *Entry Exam* ergab sich kein annähernd signifikantes Ergebnis (B = -0,35; Beta = -0,01; p = 0,909). Der individuelle Lernerfolg durch die Intervention war somit für weibliche Studierende am Ende des Beobachtungszeitraumes größer als für männliche.

In der multiplen Regression zeigte sich unter der Adjustierung für die anderen Parameter keine statistische Signifikanz mehr, jedoch deutete sich ein ähnliches Ergebnis an, dass weibliche Studierende eine höhere Prozentscore-Differenz zwischen Interventions- und Kontrollitems im *Retention Test* (B = 5,77; Beta = 0,18; p = 0,071) vorwiesen. Im *Exit Exam* fand sich kein Effekt (B = 2,08; Beta = 0,07; p = 0,521).

3.4.2 Unterschiede nach Alter

Im *Exit Exam* hatten die Studierenden der berechneten bivariaten linearen Regression zufolge mit höherem Alter eine größere Differenz zwischen Interventions- und Kontrollitems im *Exit Exam*, jedoch mit einer Irrtumswahrscheinlichkeit deutlich oberhalb des Signifikanzniveaus (B = 0,31; Beta = 0,06; p = 0,570). Im *Retention Test* war die Differenz pro Jahr höheren Alters um 1,03 Prozentpunkte kleiner bei annähernd erreichtem Signifikanzniveau

(B = -1,03; Beta = -0,19; p = 0,056). Im *Entry Exam* gab es einen geringen und nicht signifikanten Unterschied zugunsten älterer Studierender (B = 0,13; Beta = 0,02; p = 0,807).

In der multiplen Regression zeigte sich weder im *Exit Exam* noch im *Retention Test* ein Zusammenhang zwischen dem Alter und der studentischen Leistung.

3.4.3 Unterschiede nach Leistung in den Klausuren des vorherigen Semesters

In der bivariaten Regression zeigte sich mit ansteigender Leistung im vorherigen Semester eine geringere Prozentscore-Differenz zwischen Interventions- und Kontrollitems im *Exit Exam* (B = -0,32; Beta = -0,17; p = 0,091). Im *Retention Test* stellte sich dieses Ergebnis nicht dar (B = -0,03; Beta = -0,02; p = 0,869), ebenso wenig im *Entry Exam* (B = 0,13; Beta = 0,01; p = 0,946).

In der multiplen Regression wurden ebenfalls keine signifikanten Assoziationen zwischen der Klausurleistung im Vorsemester und dem Lernerfolg während der Intervention identifiziert.

3.4.4 Unterschiede nach Leistung in den Klausuren des laufenden Semesters in den Klausurfragen für Innere Medizin

Für die Ergebnisse der Modulklausuren für Innere Medizin, welche im Semester der Studie geschrieben wurden, galt, dass die leistungsstärkeren Studierenden eine signifikant geringere Differenz zwischen Interventions- und Kontrollitems im *Exit Exam* (B = -0,36; Beta = -0,21; p = 0,027) und, allerdings nicht-signifikant, im *Retention Test* (B = -0,17, Beta = -0,10; p = 0,307) zeigten. Im *Entry Exam* zeigte sich kein signifikanter Zusammenhang (B = 0,01; Beta = 0,01; p = 0,946). Der individuelle Lernerfolg durch die Intervention war demnach umso geringer, je besser die konsekutive Klausurleistung in Innere Medizin war.

In der multiplen Regression war das Ergebnis nicht statistisch signifikant (p-Werte für *Exit Exam* und *Retention Test* jeweils >0,1) Durch Herausnahme des Prozentscores der Klausuren im Sommersemester 2016 aus der multiplen Regression ließ sich der p-Wert zwar leicht senken; signifikant wurde das Ergebnis jedoch weiterhin nicht (*Exit Exam*: B = -0,34; Beta = -0,21; p = 0,057; *Retention Test*: B = -0,33; Beta = -0,20; p = 0,066). Ähnliches galt für die Herausnahme der Prozentscores der Freitextbewertungen (*Exit Exam*: B = -0,34; Beta = -0,20; p = 0,055; *Retention Test*: B = -0,30; Beta = -0,18; p = 0,081) und der Freitextlänge (*Exit Exam*: B = -0,35; Beta = -0,21; p = 0,058; *Retention Test*: B = -0,36; Beta = -0,20; p = 0,062).

3.4.5 Unterschiede nach Beantwortung der Freitextfragen

Nach Qualität der Freitextfragen

In der bivariaten Regression zeigte sich weder für das *Entry Exam*, noch für das *Exit Exam* oder den *Retention Test* ein signifikanter Zusammenhang zwischen den Freitextbewertungen und dem Lernerfolg.

In der multiplen Regression, in der unter anderem für die generelle Leistung in den Modulklausuren kontrolliert wurde, blieben diese Ergebnisse unverändert (siehe Tabelle 5).

Nach Länge der Freitextantworten

Da die Länge der Freitextantworten in Zeichen gemessen wurde, sind hier sehr kleine Koeffizienten B zu erwarten. Daher werden diese Ergebnisse angegeben als Veränderung pro zehn Zeichen Antwortlänge.

Die bivariate Regression zeigte keinen Hinweis auf eine Auswirkung der Länge der Freitextantworten auf die Prozentscore-Differenz zwischen Interventions- und Kontrollitems im *Exit Exam* (B = -0,001; Beta = -0,07; p = 0,465). Im *Retention Test* dagegen zeigte sich allenfalls eine leichte aber nicht signifikante Tendenz zu höheren Prozentscore-Differenzen bei längeren Freitexten (B = 0,002; Beta = 0,16; p = 0,109). Im *Entry Exam* ergab sich kein signifikanter Zusammenhang (B = 0,00; Beta = 0,04; p = 0,704).

In der multiplen Regression fand sich für keinen der drei Zeitpunkte ein signifikanter Zusammenhang zwischen der Freitextlänge und dem Lernerfolg.

Tabelle 5: Multiple Regression mit der Prozentscore-Differenz zwischen Interventions- und Kontrollitems im *Exit Exam* als abhängige Variable

Charakteristika	Koeffizient B	Standardisierter Koeffizient Beta	Signifikanz
Alter in Jahren	0,09	0,02	0,883
Geschlecht	2,08	0,07	0,521
Erreichte Punktzahl in Innere Medizin-Klausuren	-0,34	-0,2	0,113
Erreichte Punktzahl in Klausuren des Vorsemesters	-0,13	-0,07	0,561
Erreichte Gesamtpunktzahl der Freitextbewertungen	0,06	0,06	0,668
Gesamt-Freitextlänge	-0,01	-0,80	0,566
Gesamtmodell:	R^2 = 0,06	p = 0,349	

Tabelle 6: Multiple Regression mit der Prozentscore-Differenz zwischen Interventions- und Kontrollitems im *Retention Test* als abhängige Variable

Charakteristika	Koeffizient B	Standardisiert-er Koeffizient Beta	Signifikanz
Alter in Jahren	-0,96	-0,17	0,097

Geschlecht	5,77	0,18	0,071
Erreichte Punktzahl in Innere Medizin-Klausuren	-0,27	-0,16	0,199
Erreichte Punktzahl in Klausuren des Vorsemesters	-0,1	-0,05	0,644
Erreichte Gesamtpunktzahl der Freitextbewertungen	0,08	0,08	0,587
Gesamt-Freitextlänge	0,02	0,14	0,295
Gesamtmodell:	R^2 = 0,12	p = 0,042	

Tabelle 7: Multiple Regression mit der Prozentscore-Differenz zwischen Interventions- und Kontrollitems im *Retention Test* als abhängige Variable ohne Adjustierung für die Erreichte Punktzahl in den Klausuren des Vorsemesters

Charakteristika	Koeffizient B	Standardisiert-er Koeffizient Beta	Signifikanz
Alter in Jahren	-0,93	-0,17	0,101
Geschlecht	5,44	0,17	0,081
Erreichte Punktzahl in Innere Medizin-Klausuren	-0,33	-0,2	0,066
Erreichte Gesamtpunktzahl der Freitextbewertungen	0,09	0,09	0,542
Gesamt-Freitextlänge	0,01	0,12	0,413
Gesamtmodell:	R^2 = 0,12	p = 0,025	

3.5 Explorative Analysen

3.5.1 Leistungsunterschiede der Studiengruppen A und B

Es ergaben sich unerwartet Leistungsunterschiede zwischen den Gruppen A und B im Studienzeitraum, welche in Tabelle 8: Leistungsunterschiede zwischen Gruppe A und Gruppe B dargestellt sind. In den Modulklausuren im Wintersemester 2016/2017 hatten die Studierenden der Gruppe B durchschnittlich höhere Prozentscores als die Studierenden in Gruppe A. In den Klausurfragen des Faches Innere Medizin im Wintersemester 2016/2017 erzielten die Studierenden der Gruppe B ebenfalls ein besseres Ergebnis als die Studierenden der Gruppe A. Die Unterschiede waren statistisch signifikant.

Im *Entry Exam* sowie im *Exit Exam* gab es keinen signifikanten Leistungsunterschied zwischen den Gruppen. Im *Retention Test* dagegen erzielten die Studierenden der Gruppe B

über alle Items hinweg durchschnittlich bessere Ergebnisse als die Studierenden in Gruppe A (Prozentscores: 64,7 ± 17,7% (B) und 55,5 ± 21,5% (A), p = 0,016).

Der primäre Endpunkt der Studie – der unterschiedliche Lernerfolg in den Interventions- und Kontrollitems, gemessen in Prozentscores im *Exit Exam* – zeigt nur in Gruppe A ein signifikantes Ergebnis zugunsten der Interventionsitems. In Gruppe B gab es keinen signifikanten Unterschied zwischen Interventions- und Kontrollitems.

Tabelle 8: Leistungsunterschiede zwischen Gruppe A und Gruppe B

Prozentscores	Gesamt	Gruppe A	Gruppe B	p
Modulklausuren Sommersemester 2016	79,8 ± 7,9	79,4 ± 7,7	80,3 ± 8,1	0,522
Modulklausuren Wintersemester 2016/2017	80,2 ± 6,7	78,4 ± 7,4	81,9 ± 5,5	**0,006**
Modulklausuren Innere Medizin Wintersemester 2016/2017	76,0 ± 9,1	74,2 ± 9,9	77,8 ± 8,0	**0,039**
Entry Exam: Alle Items	27,3 ± 11,7	27,1 ± 12,7	27,4 ± 10,7	0,899
Exit Exam: Alle Items	64,0 ± 19,9	62,8 ± 21,5	65,3 ± 18,3	0,516
Retention Test: Alle Items	60,2 ± 20,1	55,5 ± 21,5	64,7 ± 17,7	**0,016**
Freitextbewertungen	57,8 ± 15,3	59,1 ± 15,0	56,6 ± 15,5	0,388

3.5.2 Prüfungsergebnisse in den Interventionsitems abhängig von der Beantwortung der Freitextfragen (Longitudinale Analyse des Antwortverhaltens)

Das Antwortverhalten in Abhängigkeit von der vorherigen Beantwortung eines Interventionsitems wird in Abbildung 6 dargestellt.

Angegeben wird für jeden Zeitpunkt, welcher Anteil der Studierenden die *Long Menu*-Fragen richtig (grün) oder falsch (rot) beantwortete bzw. die Freitextfragen mit adäquatem thematischen Bezug beantwortete (grün) oder das Thema verfehlte (rot). Dabei sind die Zahlen gemittelt über alle Interventionsitems.

Es ergibt sich ein Diagramm, in dem die Beantwortung einer Frage zu jedem Zeitpunkt in Abhängigkeit von der Beantwortung zum jeweils vorherigen Zeitpunkt dargestellt wird. Dementsprechend beziehen die Prozentzahlen sich stets auf die Gesamtmenge des vorherigen Schritts im Diagramm. Die zugehörigen absoluten Zahlen verringern sich in jedem Schritt von links nach rechts im Diagramm. Während die ersten beiden Werte (Zeitpunkt: EFS1) sich also auf die absolute Menge aller erstmalig in einem E-Fallseminar beantworteten Interventionsitems bezieht, liegt den am weitesten rechts stehenden Werten (Zeitpunkt: *Exit Exam*) ein deutlich kleinerer Absolutwert zugrunde (alle Interventionsitems, für die ein

Studierender die jeweils dargestellte Reihenfolge richtiger und/oder falscher Beantwortungen gezeigt hat). Die absoluten Werte zum Zeitpunkt *Exit Exam* reichen von 8 bis 194.

Anhand der Prozentwerte lässt sich erkennen, dass einer adäquaten (d.h. ernsthaften) Beantwortung der Freitextfragen in den meisten Fällen ein besseres Ergebnis bei erneuter Abfrage des Items folgte.

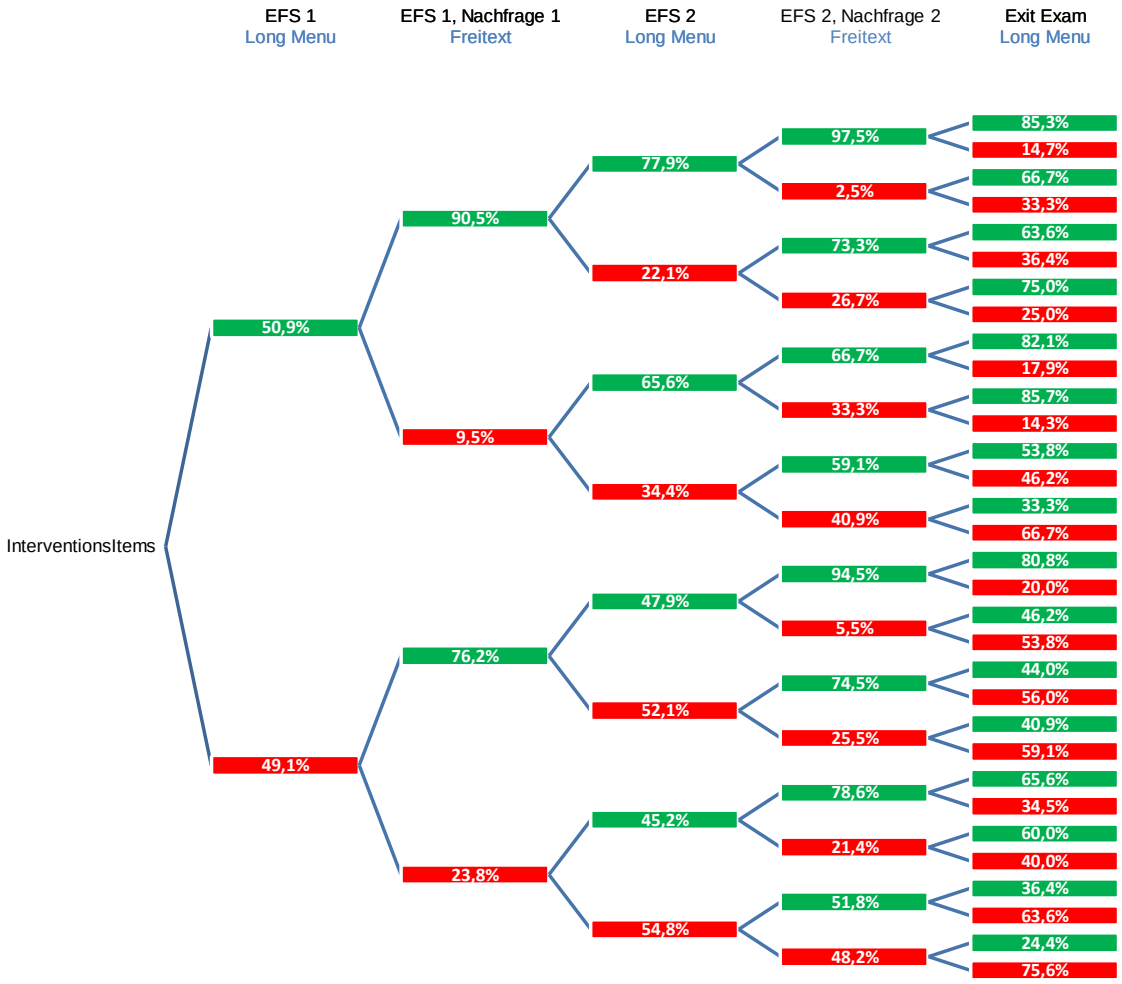


Abbildung 6: Häufigkeit von Freitextantworten der Kategorie „0“ (grob falsch oder Thema verfehlt; rot) und der Kategorien „1“ und „2“ (falsch oder unvollständig mit fachlich passendem Gedankengang bzw. richtig beantwortet; grün) im Verlauf des Studienzeitraums in Abhängigkeit von vorheriger Beantwortung der Items. EFS = E-Fallseminar. EFS 1 = erste Präsentation des Items in einem E-Fallseminar. EFS 1, Nachfrage 1 = zur ersten Präsentation zugehörige Freitextantwort. EFS 2 = zweite Präsentation des Items in einem E-Fallseminar. EFS2, Nachfrage 2 = zur zweiten Präsentation zugehörige Freitextantwort. *Exit Exam* = Präsentation des Items im *Exit Exam*.

3.5.3 Veränderung der Falschantworten in den Interventionsitems

Die Entwicklung der Häufigkeit erbrachter Falschantworten in den Interventionsitems ist in Anhang 4 dargestellt. Hier wird als Auszug daraus der Verlauf zweier beispielhafter Items gezeigt:

Tabelle 9: Antworten zum Item „Verdachtsdiagnose Perikarditis epistenocardica bei neuer ST-Hebung nach stattgehabtem Infarkt“

Item	**Häufigkeit in Prozent** (n=108)						
	Antwort	***Entry Exam***		***Exit Exam***		***Retention Test***	
Fall 1 KF4: Pericarditis epistenocardica (Interventionsitem Gruppe B)		Intervention	Kontrolle	Intervention	Kontrolle	Intervention	Kontrolle
	Richtige Antwort	9,1%	9,4%	69,1%	60,4%	60,0%	49,1%
	2 Myokardinfarkt	76,4%	69,8%	5,5%	15,1%	12,7%	18,9%
	3 Stentverschluss/Re-Infarkt	10,9%	13,2%	1,8%	5,7%	1,8%	1,9%
	4 Myokarditis	0,0%	3,8%	10,9%	5,7%	5,5%	13,2%
	5 Perikardtamponade	0,0%	0,0%	7,3%	0,0%	0,0%	0,0%
	Andere/keine Antwort	3,6%	3,8%	5,5%	13,2%	20,0%	17,0%

In Tabelle 9 und Tabelle 10 dargestellt ist die prozentuale Häufigkeit ausgewählter Antworten, darunter die korrekte und verschiedene Falschantworten. Farbig hinterlegte Spalten kennzeichnen die Falschantworten, auf welche in der Freitextfrage gesondert eingegangen wurde, von welchen die richtige Antwort also explizit abgegrenzt werden sollte. Wurde keine Falschantwort in der Freitextfrage hervorgehoben, entfällt die farbige Hinterlegung. Fett markierte Werte zeigen ein statistisch signifikant unterschiedliches Antwortverhalten zwischen Interventions- und Kontrollgruppe an.

Tabelle 10: Antworten zum Item „Erkennen einer Tachyarrhythmia absoluta im EKG“

Item	**Häufigkeit in Prozent** (n=108)						
	Antwort	***Entry Exam***		***Exit Exam***		***Retention Test***	
Fall 4 KF2: Erkennen einer Tachyarrhythmia absoluta im EKG (Interventionsitem Gruppe A)		Intervention	Kontrolle	Intervention	Kontrolle	Intervention	Kontrolle
	Richtige Antwort	30,2%	25,5%	**69,8%**	**49,1%**	45,3%	61,8%
	2 AV-Block	7,5%	21,8%	**9,4%**	**25,5%**	15,1%	16,4%
	3 Arrhythmie/Tachykardie allgemein	35,8%	32,7%	**5,7%**	**16,4%**	15,1%	12,7%
	Andere/keine Antwort	26,4%	20,0%	**15,1%**	**9,1%**	24,5%	9,1%

4 Diskussion

4.1 Einleitung

In dieser Arbeit wurde untersucht, wie die Beantwortung von Freitextfragen mit der Aufforderung, richtige und falsche Antworten zu kontrastieren, im Rahmen von formativen *Key Feature*-Prüfungen über Teilgebiete der Inneren Medizin das Lernen von *Clinical Reasoning* beeinflusst. Außerdem sollte betrachtet werden, inwiefern unterschiedliche demographische Charakteristika der Studierenden einen Einfluss auf den individuellen Lernerfolg durch eine solche Intervention haben.

Um die Forschungsfragen beantworten zu können, sollen zunächst die Ergebnisse genauer in den Blick genommen werden.

4.2 Analyse und Interpretation der Ergebnisse

4.2.1 Charakterisierung der Stichprobe

Wir betrachten eine Stichprobe aus 108 Medizinstudierenden im 3. klinischen Semester, von denen mehr als die Hälfte weiblich waren und die durchschnittlich in die Altersgruppe der Mitte-Zwanzig-Jährigen fielen. Es fällt auf, dass die Modulklausuren in Innere Medizin in dieser Kohorte durchschnittlich etwas schlechter ausgefallen sind als die Klausuren des Vorsemesters oder der Durchschnitt aller Klausuren desselben Semesters.

4.2.1.1 Teilnahme an den E-Fallseminaren

Eine Teilnahme von 80% an den E-Fallseminaren war laut Curriculum verpflichtend. Bei den Teilnahmequoten fällt auf, dass sie in den späteren Wochen tendenziell absank, vor Modulklausuren jedoch wieder stieg. Zu sehen ist das besonders gut in Woche 6 – dieses E-Fallseminar fand zeitlich kurz vor der Modulklausur des Moduls 3.1 statt.

4.2.1.2 Eigenschaften der Studiengruppen A und B

Die Studiengruppen A und B wurden im Vorfeld so gebildet, dass beide Gruppen anhand der Klausuren des Vorsemesters durchschnittlich den gleichen Leistungsstand hatten, also aus Studierenden mit vergleichbaren Klausurleistungen bestanden. Im Rahmen des Cross-Over-Designs sollte kein Vergleich zwischen den Studiengruppen A und B, sondern zwischen den Interventions- und Kontrollitems stattfinden. Zur Beantwortung der Forschungsfragen ist ein Vergleich der Studiengruppen A und B somit irrelevant. Dass sich deskriptiv dennoch Leistungsunterschiede zwischen den Gruppen im Studiensemester Wintersemester 2016/2017 zeigten, ist als Zufall zu werten. Aufgrund des Cross-Over-

Designs verfälscht dieser Umstand nicht den Vergleich zwischen Intervention und Kontrolle. Die Unterschiede zwischen den Gruppen A und B sind lediglich von explorativem Interesse (siehe unten).

4.2.2 Charakterisierung des Messinstruments: Itemkennwerte

Erwartungsgemäß war die Itemschwierigkeit im *Entry Exam* deutlich höher (zu erkennen an niedrigeren Werten) als in *Exit Exam* und *Retention Test*. Erklären lässt sich dieser Umstand dadurch, dass das *Entry Exam* vor Beginn der curricularen Lehre der abgefragten Inhalte geschrieben wurde und somit lediglich das Vorwissen der Studierenden auf dem neuen Fachgebiet abbildet. Zu den Zeitpunkten des *Exit Exam* und *Retention Test* hatten die Studierenden die abgefragten Inhalte bereits gehört beziehungsweise für die Modulklausur gelernt. Auch die niedrigeren Werte für die Itemtrennschärfe und für das Cronbachsche Alpha im *Entry Exam* lassen sich dadurch erklären. Dass die Schwierigkeit im *Retention Test* wieder größer war als im *Exit Exam*, lässt sich durch die Zeitspanne von mehreren Monaten seit Abschluss des Moduls erklären.

An den Itemkennwerten lässt sich zudem – in allen drei Prüfungen – eine große Streubreite der Werte als Hinweis auf die Heterogenität der Items erkennen. Es wird deutlich, dass einige Items den Studierenden im Durchschnitt deutlich leichter gefallen sind als andere. Die Streubreite der Itemschwierigkeiten war im *Retention Test* am größten – was einen Hinweis darauf liefert, dass der Lerneffekt in den einzelnen Items nach mehrmonatigem Follow-Up-Zeitraum heterogener war als direkt nach Abschluss der Lehre zum Zeitpunkt des *Exit Exam*. Einige Lerninhalte wurden offenbar besser behalten als andere. Die mittlere Itemtrennschärfe sowie die interne Konsistenz waren im *Retention Test* am größten - allerdings mit ausgesprochen ähnlichen Werten wie im *Exit Exam*. Daraus lässt sich schließen, dass die Fähigkeit der Prüfung, starke von schwachen Studierenden auf diesem Gebiet zu unterscheiden, nach dem mehrmonatigen Follow-Up-Zeitraum nicht abgefallen war. Insgesamt sind dies gute Hinweise darauf, dass die *Key Feature*-Prüfungen zur Abfrage des Lernerfolgs geeignet waren.

4.2.3 Forschungsfrage 1: Effekte der Intervention auf den Lernerfolg

4.2.3.1 Allgemeiner Lernerfolg

Wird zunächst der allgemeine Lernerfolg betrachtet – hier definiert als Lernerfolg in allen Items – fällt eine klare Verbesserung im Vergleich von *Entry Exam* und *Exit Exam* auf. Der allgemeine Lernerfolg lässt sich hauptsächlich durch die curriculare Lehre zwischen *Entry Exam* und *Exit Exam* erklären. Der Abfall zwischen *Exit Exam* und *Retention Test* deutet darauf hin, dass das Langzeit-*Retrieval* schlechter war als das kurzfristige *Retrieval*. Trotzdem ist noch immer – Monate nach der Lehre und den formativen *Key Feature*-Prüfungen – ein deutlicher Lernerfolg im Vergleich zum *Entry Exam* nachweisbar. Es gab somit einen kurzfristigen wie auch einen längerfristigen Lernerfolg. Diese Ergebnisse stimmen überein mit

denen aus Vorstudien zu formativen *Key Feature*-Prüfungen im Medizinstudium, die sich aus demselben Pool von Items bedienten (Raupach et al. 2016; Ludwig et al. 2018). Außerdem bestätigen sie bekannte Eigenschaften des *Testing Effect*s, der auch über Zeiträume von mehreren Monaten persistiert (siehe Kapitel 1.5; Nungester und Duchastel 1982). Dass die Ergebnisse im *Retention Test* schlechter ausfielen als im *Exit Exam*, scheint auf den ersten Blick nicht zur Literatur zum *Testing Effect* zu passen: hier wurde beschrieben, dass sich der *Testing Effect* mehr im langfristigen als im kurzfristigen Lernerfolg zeigt (Roediger und Karpicke 2006b). Allerdings handelte es sich bei den zitierten Grundlagenexperimenten meist um deutlich kürzere Zeiträume von Tagen bis Wochen und um Inhalte und Methoden fernab des Medizinstudiums. Im Vergleich zu anderen Studien über Test-Enhanced-Learning im Medizinstudium (Larsen et al. 2009; Schmidmaier et al. 2011; Larsen et al. 2013) fiel der Verlust des Lernerfolgs im mehrmonatigen Intervall geringer aus. Es sollte jedoch nicht unerwähnt bleiben, dass der lange Follow-Up-Zeitraum viele mögliche, nicht kontrollierte Confounder mit sich bringt (Raupach und Schuelper 2018).

4.2.3.2 Spezieller Lernerfolg in Interventionsitems

Betrachtet man die Ergebnisse von Interventions- und Kontrollitems getrennt, zeigt sich im *Exit Exam* ein größerer Lernerfolg in Interventionsitems als in den Kontrollitems. Trotz nahe beieinander liegender durchschnittlicher Prozentscores ist der Unterschied statistisch signifikant. Die Effektstärke von d = 0,16 fällt eher gering aus. Insgesamt zeigt sich in diesen Zahlen eine überschaubare, aber signifikant nachweisbare Überlegenheit der Interventionsitems, *Clinical Reasoning*-Kompetenz zu vermitteln nach einer Lernphase mit wiederholten formativen *Key Feature*-Prüfungen innerhalb eines Semesters. Die Aufforderung, Falschantworten zu reflektieren und die richtige Lösung zu begründen, führte also zu einer besseren kurzfristigen Retention der Lerninhalte als die reine Prüfung durch *Key Feature*-Items im *Long Menu*-Format. Die bereits als wirksam erwiesene Methode, formative *Key Feature*-Prüfungen zum Lernen von *Clinical Reasoning* einzusetzen (Raupach et al. 2016), konnte somit durch die Implementierung der Freitextfragen mit besagter Reflexion weiter verbessert werden.

Im *Entry Exam* zeigte sich, wie erwartet, noch kein signifikanter Unterschied zwischen Interventions- und Kontrollitems. Im *Retention Test*, ca. vier Monate nach Durchführung des *Exit Exams*, konnte die Überlegenheit der Interventionsitems nicht mehr nachgewiesen werden, obwohl der allgemeine Lernerfolg nur geringfügig gesunken war (s.o.). Der Effekt der Freitextfragen schien im mehrmonatigen Intervall verschwunden zu sein. Dabei muss jedoch beachtet werden, dass bereits der initiale Effekt zum Zeitpunkt des *Exit Exams* eher gering war. Bei noch zusätzlich vorliegendem Abfall des allgemeinen Lernerfolgs im Intervall war im *Retention Test* daher kein großer Unterschied mehr zu erwarten. Beachtet werden muss auch die bereits erwähnte Tatsache, dass in diesem langen Zeitraum vermutlich schwierig zu kontrollierende interferierende Faktoren vorlagen (Raupach und Schuelper 2018) und dass davon auszugehen ist, dass eine weitere Auseinandersetzung mit oder An-

wendung des gelernten Wissens in diesem Zeitintervall nicht stattfand, während bereits neue Themen gelernt wurden. Für eine bessere Beurteilung, ob eine Überlegenheit der Interventionsitems auch nach mehrmonatigem Intervall noch besteht, wären ein größerer initialer Effekt, kontrolliertere Bedingungen im Follow-Up-Zeitraum und eine größere Probandenzahl wünschenswert. Die vorliegenden Ergebnisse können nur zeigen, dass direkt nach der Intervention der deutlichste Effekt bestand.

Im Vergleich zu den Kontrollitems wurden die richtigen Ergebnisse der Interventionsitems besser behalten – das *Retrieval* konnte durch die Elaboration also verbessert werden. Ein möglicher zugrundeliegender Mechanismus kann in den Eigenschaften des *Testing Effect*s gefunden werden. In der Literatur wird beschrieben, dass mühsameres, tieferes *Retrieval* von Informationen den *Testing Effect* verstärkt (Roediger und Karpicke 2006b). Es wurde bereits in anderen Studien gezeigt, dass aufwändigere Fragenformate (die mit einer Produktion von Antworten einhergehen) einen größeren *Testing Effect* bewirken als *Multiple Choice Question* (MCQ)-Tests, die nur mit einem Wiedererkennen der Antworten einhergehen (Roediger und Karpicke 2006b; Larsen et al. 2008). In dieser Studie wurden die Freitextantworten zwar nicht mit *Multiple Choice*-Fragen verglichen, sondern mit aufwändigeren *Long Menu*-Fragen (die eine gewisse eigene Produktion der Antwort voraussetzen), jedoch ist davon auszugehen, dass die freie schriftliche Abgrenzung und Verteidigung von richtigen und falschen Aussagen – wie in den Interventionsitems geschehen – mit einem mühsameren und tieferen *Retrieval* einhergeht als die Beantwortung einer Frage mit einem einzigen, stichwortartigen Begriff. Dass durch das Erklären eines Sachverhaltes das Lernen allgemein erleichtert wird, wurde bereits in der Einleitung erwähnt unter Verwendung des Begriffs generative Lernstrategie (Hasselhorn und Gold 2009; Weinstein et al. 2011). Hierbei spielt die Elaboration als kognitive Strukturierung des Lernmaterials eine Rolle (Slavin 1996). Dass der Lernerfolg in dieser Studie durch die Freitextfragen in den Interventionsitems gesteigert wurde, stimmt mit diesen Prinzipien überein.

4.2.3.3 Mögliche Gründe für einen eingeschränkten Lernerfolg

Anhand theoretischer Überlegungen könnte ein größerer Unterschied zwischen dem Lernerfolg in Interventions- und Kontrollitems erwartet werden als der hier nachgewiesene. Was könnte den Effekt geschmälert haben? Zum einen kommen organisatorische Faktoren infrage. Die Teilnahmequote an den E-Fallseminaren war meist nicht 100% – Studierende verpassten also einzelne *Key Feature*-Prüfungen in der Lernphase und dadurch einzelne Items bzw. Wiederholungen von Items. Diese spielten im Sinne des repetitiven Testens als Lernansatz eine wichtige Rolle in der Festigung des Wissens. Hierdurch ist zwar in erster Linie ein geringerer allgemeiner Lernerfolg zu erwarten, aber es kann im Einzelfall nicht ausgeschlossen werden, dass mehr Interventions- als Kontrollitems verpasst wurden. Weiterhin war die Bearbeitung der Freitextfragen im Grunde freiwillig. Es musste zwar zwingend etwas in das Freitextfeld eingegeben werden, jedoch genügte ein einzelnes Zeichen. Manche Studierende gaben elaborierte Freitextantworten, andere schrieben jedoch nur

knappe Antworten, die teils kürzer als ein Satz waren. Wiederholt kam es auch vor, dass Studierende gar keine inhaltliche Antwort produzierten, sondern beispielsweise „keine Ahnung“ schrieben. Dadurch konnte ein geringerer Lerneffekt entstehen.

Der vermutlich größte Nachteil an den Freitextfragen war, dass die Nachfrage nicht auf den individuellen Bearbeiter zugeschnitten war, sondern für alle Interventionsitems gleich. Das bedeutet, dass die Studierenden mit Falschantworten konfrontiert wurden, die nicht ihre eigenen waren. Es war zwar möglich, dass sie die jeweilige Falschantwort gegeben hatten – anhand der Daten aus früheren Semestern wurden schließlich die häufigsten Falschantworten ermittelt und aufgegriffen –, aber abgesehen davon war die Nachfrage vollkommen unabhängig von der individuellen Antwort. In dem Fall, dass ein Studierender das Item zuvor richtig beantwortet hatte, konnte die Abgrenzung von möglichen Falschantworten immerhin das korrekte Wissen konsolidieren. Hatte ein Studierender dagegen eine Falschantwort gegeben, die in der Nachfrage nicht vorkam, wurde der eigene Gedankengang nicht aufgegriffen. Daher war die metakognitive Komponente der Freitextfragen begrenzt. Es ist davon auszugehen, dass die Elaboration der Lösungen unabhängig davon ihren Effekt hat, aber womöglich wäre dieser Effekt durch die Reflektion eigener Gedankengänge größer. In der Literatur über das Lernen von *Clinical Reasoning* wird vorgeschlagen, dem Lernenden idealerweise sofortiges Feedback mit Diskussion der eigenen Fehler zu bieten (Kassirer 2010; Levine und Bleakley 2013).

Darüber hinaus war auch die Nutzung des Feedbacks freiwillig und nicht individualisiert. Der ausführliche Feedbacktext konnte nach jeder Frage durchgelesen oder ignoriert werden. Hinzu kommt die Tatsache, dass das Feedback nicht individualisiert war, sondern stets ein standardisierter Feedbacktext gegeben wurde. In einer anderen Studie wurde gezeigt, dass ein vom Lernenden selbst anpassbares Feedback bei der Bearbeitung von Fällen mit Falschbeispielen mehr *Clinical Reasoning*-Lernerfolg brachte als selbst produzierte Erklärungen (Heitzmann et al. 2015). Die Bedeutung von ausreichend tiefgehendem Feedback zur Kompensation der erhöhten Komplexität bei der Bearbeitung von Falschbeispielen wurde auch durch eine Studie von Stark et al. gezeigt (Stark et al. 2011). Bei dem in den *Key Feature*-Prüfungen verwendeten Feedback handelte es sich um ein elaboriertes Feedback, welches den Studierenden Zusammenhänge und Herleitungen erklärte und damit eine recht detaillierte Unterstützung bot. Es gab jedoch kein individuelles Feedback, das auf die Antworten der einzelnen Studierenden zugeschnitten war.

An den Ergebnissen in den einzelnen Interventionsitems ist zu erkennen, dass der durchschnittliche Lernzuwachs im Vergleich zwischen den Items sehr unterschiedlich ausfiel. Einige Items zeigten einen besonders großen Effekt der Intervention, andere gar keinen. Nicht alle Interventionsitems trugen daher gleichermaßen zum Gesamteffekt bei. Als Erklärung ist denkbar, dass die Intervention in dieser Form für manche Lerninhalte besser geeignet war als für andere. Es ist bereits bekannt, dass *Clinical Reasoning*-Kompetenz in hohem Maße kontext- beziehungsweise fallspezifisch ist (Elstein et al. 1978) und daher

wird empfohlen, beim Lernen und Prüfen von *Clinical Reasoning* eine möglichst große Breite an Fällen zu erfassen (Eva 2005). Die hier beobachtete Heterogenität der Items könnte dafür eine Bestätigung sein. Warum im Einzelnen manche Items weniger empfänglich für die Intervention waren als andere, geht aus den Daten nicht hervor.

Zu möglichen Confoundern, die mit der Messung des eigentlichen Lerneffekts durch die Elaboration interferiert haben könnten, gehört die Bearbeitungszeit. Wer für die Interventionsitems Freitextantworten schrieb, beschäftigte sich mutmaßlich länger mit dem Lerninhalt, als es bei der Bearbeitung von Kontrollitems der Fall war. Die Bearbeitungszeiten der einzelnen Items wurden in dieser Studie nicht aufgezeichnet, sodass es sich hierbei nur um eine Vermutung handelt. Durch die längere Exposition könnte das Wissen sich besser gefestigt haben. Dies muss jedoch nicht zwangsläufig ein Nachteil sein: auch ein durch längere Bearbeitungszeit erzielter zusätzlicher Lernerfolg ist ein Lernerfolg.

Ebenso können Überlegungen zur Motivation der Studierenden herangezogen werden. Die Motivation wurde in dieser Studie nicht eingehend gemessen, sondern kann nur abgeschätzt werden anhand von Parametern wie Qualität und Quantität der Freitexte sowie der Teilnahmequoten. Dass letztere im Laufe des Interventionszeitraums leicht absanken und erst kurz vor summativen Prüfungen wieder anstiegen, wurde bereits erwähnt. Viele grob falsche, vom Thema abweichende oder schlichtweg verweigerte Freitextantworten deuteten auf mangelnde Sorgfalt des Antwortgebers hin. Aus einer dieser Freitextantworten ging hervor, dass die zeitliche Lage der E-Fallseminare in der Mittagszeit nicht zum Abrufen kognitiver Bestleistungen führte. Selbstverständlich handelt es sich dabei nur um eine einzelne subjektive Aussage. Dass die Motivation der Studierenden, alle Fragen bestmöglich zu beantworten, zumindest nicht bei allen und nicht immer gleich hoch war, ist jedoch gut denkbar. Letztendlich sind Medizinstudierende aufgrund des Aufbaus ihres Studiums ein hohes Pensum summativer Prüfungen gewöhnt, jedoch nicht das Lernen mittels formativer Prüfungen. Die Fokussierung auf die nächste anstehende Modulklausur könnte dazu führen, dass anderen nicht-summativen Prüfungsformaten eine geringere Priorität zugeschrieben wird. *Assessment drives learning* bedeutet auch, dass für das Bestehen der Klausur gelernt wird – und nicht in erster Linie für die spätere Praxis. Sofern eine Klausur eher Faktenwissen abprüft als *Clinical Reasoning*, besteht für die Studierenden eine geringere Motivation, *Clinical Reasoning* zu lernen. Insofern muss auf *Constructive Alignment* geachtet werden: die summative Prüfung sollte genau das prüfen, was gelernt werden soll (Biggs 1996; Daniel et al. 2019). Die an dieser Studie teilnehmende Kohorte von Studierenden wurde in der Modulklausur des Moduls 3.1 mit einer Mischung aus *Multiple Choice*-Fragen und *Key Feature*-Fragen geprüft. Somit war zumindest ein Teil der summativen Prüfung auf das Lernziel *Clinical Reasoning* und die Lernmethode der *Key Feature*-Prüfungen ausgerichtet. Jedoch traf dies nicht für die übrigen Modulklausuren zu, also für die Art und Weise, wie die Studierenden gewohnt waren, geprüft zu werden. Hier herrschten weiterhin die traditionellen *Multiple Choice*-Fragen vor. Der Anreiz, fallspezifisches und praxisnahes *Clinical Reasoning* zu lernen, kann damit als geringer angenommen werden als der Anreiz, *Multiple Choice*-fähiges

Faktenwissen zu lernen. Eine andere Erklärung, warum die Freitextantworten teilweise nur halbherzig bearbeitet wurden, kann bei VanLehn gefunden werden (VanLehn 1999). Hier wird beschrieben, dass *Self-explanation* – das Generieren von eigenen Erklärungen durch den Lernenden zur Vertiefung des Gelernten, zum Beispiel durch Herleitung anhand zugrundeliegender Mechanismen (Chamberland et al. 2020) – mehr Aufwand erfordert und eher langfristigen als kurzfristigen Lernerfolg bewirkt. Dass Studierende diesen Aufwand oft vermieden, kann an ihrer stärkeren Orientierung an kurzfristigem Lernerfolg liegen. Gleiches gilt für *Test-Enhanced Learning* im Allgemeinen: da es mühsamer ist als Lernen durch Wiederholung (wobei es ebenfalls stärkere Auswirkungen auf den langfristigen als auf den kurzfristigen Lernerfolg hat), neigen Lernende eher dazu, sich dagegen zu entscheiden (Roediger und Karpicke 2006b). Obwohl die Motivation der Studierenden in dieser Studie nicht gemessen wurde, gibt es also plausible Erklärungen dafür, warum sie vermindert gewesen sein könnte.

4.2.3.4 Gruppe A als Treiber hinter dem Lerneffekt

Zuletzt fällt auf, dass der Treiber hinten dem gemessenen Lerneffekt durch die Interventionsitems die Studiengruppe A war. Nur in Gruppe A zeigte sich ein signifikanter Unterschied zwischen Interventions- und Kontrollitems im *Exit Exam* ($p = 0{,}008$), in Gruppe B war der Unterschied nicht annähernd signifikant ($p = 0{,}659$) – obwohl es in der Exposition zur Intervention formal keine Unterschiede zu Gruppe A gab. Zwei Unterschiede zwischen den beiden Gruppen können mögliche Erklärungen liefern: zum einen hatten sie unterschiedliche Interventionsitems bekommen. Diese wurden zwar anhand der Ergebnisse des *Entry Exams* so auf die Gruppen verteilt, dass die Itemschwierigkeiten sich nicht unterschieden. Aber ein unerwarteter Unterschied während und nach der Intervention ist nicht ausgeschlossen. Es wurde bereits diskutiert, dass *Clinical Reasoning* einer hohen Kontextspezifität unterliegt. Es könnte sein, dass die der Gruppe A zugeordneten Interventionsitems besser auf die Intervention angesprochen haben, weil sie besser für diese Art des Lernens geeignet waren. Dafür ergeben die Daten jedoch keine direkten Hinweise, daher kann darüber keine klare Aussage getroffen werden.

Zum anderen sieht man einen Leistungsunterschied zwischen den Studierenden in Gruppe A und in Gruppe B in den Klausuren des Studiensemesters: sowohl in allen Klausuren des Studiensemesters zusammen als auch in den Modulklausuren der Inneren Medizin waren die Studierenden der Gruppe B signifikant besser. Die Studiengruppen A und B wurden im Vorfeld zwar so gebildet, dass beide Gruppen anhand der Klausuren des Vorsemesters durchschnittlich den gleichen Leistungsstand hatten, also aus Studierenden mit vergleichbaren Klausurleistungen bestanden. Im Rahmen des Cross-Over-Designs sollte kein Vergleich zwischen den Studiengruppen A und B, sondern nur zwischen den Interventions- und Kontrollitems stattfinden. Dass sich deskriptiv dennoch Leistungsunterschiede zwischen den Gruppen im Studiensemester Wintersemester 2016/2017 zeigten, ist als unvorhersehbarer Zufall zu werten. Aufgrund des Cross-Over-Designs verfälscht dieser Um-

stand nicht den Vergleich zwischen Intervention und Kontrolle. Zur Beantwortung der Forschungsfrage 1 ist ein Vergleich der Studiengruppen A und B irrelevant. Die Unterschiede zwischen den Studiengruppen sind jedoch von explorativem Interesse – um aufgrund der zufällig aufgetretenen gegenläufigen Leistungsunterschiede mehr über den Effekt der Intervention herauszufinden. Ob es hier einen Zusammenhang mit der allgemeinen Leistung gibt, wird anhand von Forschungsfrage 2 untersucht.

4.2.4 Forschungsfrage 2: Ergebnisunterschiede zwischen den Studierenden unterschiedlicher Charakteristika

4.2.4.1 Unterschiede nach Geschlecht

In der bivariaten Regression zeigt sich, dass weibliche Studierende mehr von der Intervention profitiert haben als männliche, insbesondere im längeren Zeitraum bis zum *Retention Test*. Bei den weiblichen Studierenden zeigen sich ein signifikant stärkerer Langzeiteffekt und ein statistisch nicht signifikanter Hinweis auf einen stärkeren Kurzzeiteffekt durch die Intervention. Wird auf weitere Eigenschaften wie Studienleistungen und Freitextparameter kontrolliert, wird dieser Effekt statistisch schwächer. Das bedeutet, dass es einen Zusammenhang zwischen dem weiblichen Geschlecht und den anderen Parametern gibt. Am stärksten ist der Zusammenhang zwischen dem Geschlecht und der Leistung in den Klausurfragen für Innere Medizin: die weiblichen Studierenden schnitten tendenziell etwas schlechter in Innere Medizin ab als die männlichen Studierenden (T-Test, mittlerer Prozentscore in Innere Medizin, männlich: 77,0 ± 8,5; weiblich: 75,5 ± 9,5; $p = 0{,}401$).

Der hier gemessene Effekt des Geschlechts ist also nicht allein auf das Geschlecht selbst zurückzuführen, sondern mindestens anteilig auf die unterschiedlichen Leistungen in der Inneren Medizin von männlichen und weiblichen Studierenden.

4.2.4.2 Unterschiede nach Alter

Die Ergebnisse der bivariaten und der multiplen Regression deuten beide auf einen nennenswerten Effekt des Alters im *Retention Test* hin. Ältere Studierende scheinen das Gelernte über einen längeren Zeitraum hinweg schneller wieder vergessen zu haben als jüngere Studierende. Neben einer reinen neurobiologischen Alterserscheinung kann als Ursache auch diskutiert werden, dass im Medizinstudium ältere Studierende tendenziell eine schlechtere Abiturdurchschnittsnote hatten als jüngere Studierende und damit länger auf einen Studienplatz warten mussten. Möglicherweise spielt dies als Hinweis auf eine geringere allgemeine Leistungs- oder Lernfähigkeit eine Rolle. Höheres Alter korrelierte außerdem stark mit einer schlechteren Klausurleistung in Innere Medizin und mit einer schlechteren Freitextqualität.

4.2.4.3 Unterschiede nach Leistung in den Klausuren des vorherigen Semesters

In der bivariaten und multiplen Regression zeigen sich keine besonders aussagekräftigen Ergebnisse. Der leicht geringere Effekt der Intervention bei leistungsstärkeren Studierenden im *Exit Exam* in der bivariaten Regression erreicht nicht das statistische Signifikanzniveau, befindet sich aber zumindest in der Nähe und kann daher als Hinweis wahrgenommen werden. Mit einer größeren Stichprobe könnte der Effekt signifikant werden. In der multiplen Regression wird er deutlich schwächer, was wieder damit zu tun haben könnte, dass die Vorleistungen mit den anderen Parametern in Zusammenhang stehen. Insgesamt ergibt sich hier ein erster, wenn auch schwacher, Hinweis darauf, dass leistungsstärkere Studierende weniger von der Intervention profitiert haben.

4.2.4.4 Unterschiede nach Leistung in den Klausuren des laufenden Semesters in den Klausurfragen für Innere Medizin

Die Ergebnisse der bivariaten Regression und, in etwas geringerem Maße, auch der multiplen Regression, lassen vermuten, dass es einen gegenläufigen Zusammenhang zwischen Klausurleistungen in der Inneren Medizin und einem Effekt durch die Intervention gab: je besser die generelle Leistung in Innere Medizin, desto weniger hatten die Studierenden von der Intervention profitiert. Dieses Ergebnis erreicht in der multiplen Regression zwar nicht ganz das statistische Signifikanzniveau, aber es ist möglich, dass es bei einer größeren Stichprobe signifikant werden würde. Insbesondere, wenn nicht mehr für die Vorleistungen (Klausurergebnisse des vorherigen Semesters) adjustiert wird, verdeutlicht sich dieses Ergebnis und gewinnt an Aussagekraft. Das bedeutet, dass nicht nur die alleinige Leistung in Innere Medizin, sondern mehr die allgemeine Studienleistung einen Unterschied macht. Generell bessere Studierende scheinen von der Intervention weniger profitiert zu haben als generell weniger starke Studierende. Sowohl die Vorleistungen, als auch die Freitextlänge und –qualität korrelieren mit der Leistung in Innere Medizin. Das erscheint plausibel, da es leistungsstärkeren Studierenden leichter gefallen sein sollte, gute und ausführliche Freitextantworten zu geben. Die Literatur bestätigt zudem: "Good students preferred deep reasoning more than poor did" (VanLehn 1999).

4.2.4.5 Unterschiede nach Beantwortung der Freitextfragen

Ohne auf andere Parameter zu kontrollieren, zeigt sich zunächst ein negativer Zusammenhang zwischen der Freitextqualität und dem Effekt der Intervention. Als Confounder besteht hier jedoch die Leistung in Innere Medizin, welche mit der Qualität der Freitextfragen korreliert. Wenn dafür adjustiert wird, verschwindet der negative Zusammenhang. Anhand des Ergebnisses der multiplen Regression lässt sich nicht beweisen, dass gute Freitexte besser für das Outcome der Intervention sind, aber zumindest zeichnet sich auch nicht das Gegenteil ab. Für den Langzeiteffekt der Intervention scheinen bessere Freitexte - abhängig von ihrer Länge – eine größere Rolle zu spielen.

Betrachtet man die Freitextlänge, scheinen längere Texte unabhängig von ihrer Qualität im *Exit Exam* eher eine negative Auswirkung auf den Effekt der Intervention zu haben und im *Retention Test* eher eine positive. Dabei sind die Ergebnisse jedoch vorsichtig zu interpretieren, da die Effektstärken gering und noch recht weit entfernt vom Signifikanzniveau sind. Einzig die Ergebnisse des *Retention Tests* sind deutlicher zugunsten von längeren Freitexten – aber nicht unabhängig von ihrer Qualität.

Insgesamt zeichnet sich ab, dass längere und bessere Freitextantworten zumindest für den Langzeiterfolg günstiger sind. Von einem sehr ausgeprägten Effekt kann anhand dieser Daten aber nicht gesprochen werden. Es fällt auf, dass die Auswirkung der Freitexte geringer wird, wenn auf die Studienleistungen kontrolliert wird, was erneut für die Korrelation zwischen den Studienleistungen und besseren bzw. längeren Freitexten spricht. Treibender Faktor scheint also noch immer die allgemeine Leistung der Studierenden zu sein.

4.2.4.6 Zusammenfassung der Ergebnisse zu Forschungsfrage 2

Die deutlichste Aussage, die sich anhand der präsentierten Ergebnisse abzeichnet, ist, dass Studierende mit besseren Studienleistungen – nicht nur auf dem Gebiet Innere Medizin, sondern auch in den anderen Klausuren – weniger von der Intervention profitierten. Leistungsschwächere Studierende konnten dagegen mehr profitieren. Mit Ausnahme des Alters – welches gleichzeitig mit schlechteren Studienleistungen und einem geringeren Langzeiteffekt der Intervention einherging – können alle Regressionsparameter diesen Befund unterstützen. Dazu passt auch die Beobachtung, dass sich bei den zufällig leistungsschwächeren Studierenden der Gruppe A ein größerer Effekt der Intervention zeigte als in der stärkeren Gruppe B. Plausibilität gewinnt dieses Ergebnis, wenn man bedenkt, dass leistungsstärkere Studierende sich durch die stärkere kognitive Auseinandersetzung mit den Interventionsitems weniger verbessern konnten, weil sie schlichtweg von vorneherein über eine gute Kompetenz verfügten und daher weniger Raum zur Verbesserung hatten. Ihre Antworten waren mit höherer Wahrscheinlichkeit von vorneherein richtig, sodass die reflektive Komponente der Freitextfragen ihren eigenen Gedankengang nicht verbessern, sondern allenfalls bewusst machen konnte. Schwächeren Studierenden dagegen half offenbar die Auseinandersetzung mit Falschantworten, die möglicherweise ihre eigenen Fehlüberzeugungen repräsentierten, ihre Konzepte und Denkmuster zu prüfen und zu korrigieren.

4.2.5 Explorative Analysen

4.2.5.1 Prüfungsergebnisse in den Interventionsitems abhängig von der Beantwortung der Freitextfragen (Longitudinale Analyse des Antwortverhaltens)

Das Baumdiagramm zeigt, dass Studierende, die sich mehr Mühe bei der Beantwortung der Freitextfragen gaben, hinterher bessere Ergebnisse in diesen Items zeigten, also mehr von der Intervention profitierten. Wurden die Freitextfragen nicht ernsthaft oder grob falsch beantwortet, ergab sich tendenziell weniger Verbesserung und ein weniger gutes Ergebnis

in diesen Items. Dies spricht für eine Effektivität der Intervention und untermauert die zugrundeliegende Theorie, dass aufwändigere Fragen für tieferes *Retrieval* und damit für einen stärkeren *Testing Effect* und eine bessere kognitive Verankerung des Wissens sorgen (Roediger und Karpicke 2006b, s.o.)

4.2.5.2 Veränderung der Falschantworten in den Interventionsitems

Die Auflistung der Interventionsitems und wie die Häufigkeiten der Falschantworten sich im Laufe der Studie veränderten, zeigt in erster Linie, wie heterogen die Items ausfielen. Das ist zumindest insofern mit unserem Vorwissen über *Clinical Reasoning* konsistent, als dass in der Literatur häufig erwähnt wird, wie kontext- und fallspezifisch *Clinical Reasoning*-Kompetenz ist. Manche Items fielen einer großen Zahl von Studierenden leichter als andere, was darauf schließen lässt, dass die Kohorte sich darin ähnelte, in welchen Fällen sie stärker im *Clinical Reasoning* waren. Das ist dadurch zu erklären, dass es sich um eine Gruppe Studierender mit überwiegend gleichem Lernstand bei gleichem Fortschritt im Medizinstudium handelte. Natürlich kommen als Erklärung für die Heterogenität der Itemresultate auch eine Reihe Confounder in Frage: zum Beispiel die Qualität der curricularen Lehre auf den verschiedenen Themengebieten, die aufgrund der unterschiedlichen Verantwortlichen, organisatorischen Gegebenheiten (Ort, Zeit, Format der Lehre und Lehrausfälle) und Dozenten durchaus Schwankungen unterliegen kann, aber auch die Zeit und Intensität, die die Studierenden zum Lernen der verschiedenen Themengebiete aufgebracht hatten. Trotz größter Sorgfalt bei der Erstellung der Items ist es möglich, dass nicht alle Items gleichermaßen den Ansprüchen hinsichtlich *Constructive Alignment* für *Clinical Reasoning* gerecht wurden. Und schließlich zeigen sich die bereits präsentierten Itemkennwerte unterschiedlich mit breiter Verteilung.

4.2.6 Zusammenfassung der Ergebnisse mit Beantwortung der Forschungsfragen

Forschungsfrage 1 lautete: „Wie wirkt sich die Aufforderung zur Elaboration und Kontrastierung von Falschantworten im Rahmen formativer *Key Feature*-Prüfungen zusätzlich zur Darbietung eines elaborierten Feedback-Texts auf den Erwerb und die langfristige Retention von *Clinical Reasoning* aus?“ Mit dieser Studie konnte gezeigt werden, dass *Key Feature*-Items, in denen eine solche Aufforderung vorkam, zu einem größeren kurzfristigen Lernerfolg führten als Items, in denen nur ein elaboriertes Feedback gezeigt wurde. Die Elaboration und Kontrastierung von Falschantworten führte somit zu einem besseren Erwerb von *Clinical Reasoning*.

Forschungsfrage 2 lautete: „Inwiefern unterscheiden sich die Studierenden hinsichtlich der Effekte der Präsentationsform auf ihren Lernprozess?“ Die nicht-adjustierten Ergebnisse der linearen Regression zeigen an, dass leistungsschwächere Studierende einen größeren Effekt durch die Intervention erzielen könnten als leistungsstärkere und damit mehr von dieser Präsentationsform profitieren könnten. Weitere Studien sind notwendig, um diesen Zusammenhang näher zu untersuchen.

4.3 Stärken und Limitationen

4.3.1 Limitationen

Anhand dieser Studie können nur Aussagen über das getroffen werden, was aus der konkreten Umsetzung vonseiten der Lehrenden und Studierenden entstand – was zum Beispiel damit zusammenhing, wie die Studierenden das Format annahmen –, weniger über das theoretische Potential der Intervention. Zwischen theoretischer Konzeption und Umsetzung in der Praxis lagen viele Schritte, die auch im Rahmen eines Experimentaldesigns nicht bis ins Letzte kontrolliert werden konnten.

Die Aufforderung, Falschantworten zu diskutieren und richtige Antworten abzugrenzen, sollte zu tieferem *Retrieval* und verstärkter kognitiver Elaboration führen. Wie tief die Studierenden sich im Endeffekt mit den Fällen beschäftigten, konnte nicht kontrolliert, sondern nur angestoßen werden. Der gesamte Lerneffekt ist dadurch limitiert, wie ernsthaft und engagiert die Items, insbesondere die Interventionsitems, bearbeitet wurden. Nur durch eine ausreichend tiefere Beschäftigung mit den Inhalten der Interventionsitems ist ein deutlicher Unterschied zu den Kontrollitems zu erwarten. Zu einer geringeren Intensität der Motivation und Konzentration der Studierenden könnte der formative Charakter der *Key Feature*-Prüfungen beigetragen haben. Für die Bearbeitung der Freitexte gab es zwar einen Anreiz in Form eines Büchergutscheins, jedoch fehlten kurzfristige externe Anreize. Die Medizinstudierenden im Regelstudiengang an der Universitätsmedizin Göttingen waren außerdem unerfahren darin, schriftliche Freitextantworten auf klinische Fragen zu geben. Durch die eher ungeübte Ausdrucksweise, teils fehlende Satzstrukturen und damit möglicherweise weniger tiefes Argumentieren wurde, unabhängig von der Motivation, eventuell kein tiefes Dekodieren und damit kein voller Effekt erreicht. Die Studierenden wurden zudem in den Freitextfragen dazu aufgefordert, stichwortartig und kurz zu antworten. Für den Effekt durch die Intervention kommen weitere Confounder infrage, wie bereits oben beschrieben, insbesondere die längere Bearbeitungszeit der Interventionsitems.

Als limitierender Faktor für den Lerneffekt gilt außerdem, dass die Freitextfragen unabhängig von der individuellen Beantwortung des *Key Feature*s waren und damit nicht auf den Lernprozess des einzelnen Studierenden zugeschnitten. Die Studierenden wurden nicht individuell mit ihren Fehlern konfrontiert – daher wurden nicht unbedingt eigene Fehler erklärt, sondern lediglich allgemeine häufige Fehler. Dadurch konnte nicht bei allen Studierenden der eigene Gedankengang aufgegriffen und die eigene *Clinical Reasoning*-Kompetenz gestärkt werden. Über das Lernen aus Fehlern ist jedoch bekannt, dass "[g]erade spezifisch eigene oder typische Fehler […], wenn sie aufgezeigt werden, eine Möglichkeit [bieten], die Lernbemühungen auf den individuellen Lerner und seine Bedürfnisse abzustimmen" (Eva 2005). Noch dazu bestand so das Risiko, negative *Cueing*-Effekte (Hinweisreizeffekte) bei denen auszulösen, die das *Key Feature* richtig beantwortet hatten und dann aufgefordert wurden, sich mich Falschantworten auseinanderzusetzen.

Die Ergebnisse der Studie sind auch durch die verwendeten Items limitiert. In diesem Falle wurden 30 *Key Feature* Items aus dem Gebiet der Inneren Medizin, davon größtenteils aus der Kardiologie, aber auch aus der Nephrologie, für die Intervention ausgewählt. Dadurch ist selbstverständlich weder die gesamte Bandbreite des Fachgebiets, noch des *Clinical Reasoning* insgesamt repräsentiert. Es konnten nur Teilgebiete abgebildet werden.

Hinzu kommt, dass der Erwerb von *Clinical Reasoning* ein langfristiger Prozess ist und dessen Art und Funktionsweise im Zuge wachsender klinischer Erfahrung veränderlich ist, wie oben beschrieben wurde (Grant und Marsden 1988; Schmidt et al. 1990; Schmidt und Boshuizen 1993; Rikers et al. 2000). Es wurde in dieser Studie das *Clinical Reasoning* von Medizinstudierenden im Anfängerstadium abgebildet. Daher muss in Betracht gezogen werden, dass die Ergebnisse nicht unbedingt auf das Lernen von *Clinical Reasoning* in allen Erfahrungsstufen generalisiert werden können.

Die anhand der Ergebnisse getroffenen Aussagen sind auf die Kohorte der teilnehmenden Studierenden limitiert. Dazu gehört nicht nur der Leistungsstand der Studierenden im dritten klinischen Semester, sondern auch die begrenzte Anzahl der Studierenden, die demographische Verteilung beispielsweise auf Alter und Geschlecht und die spezifischen curricularen Gegebenheiten an der Universitätsmedizin Göttingen im Wintersemester 2016/2017.

Weiterhin kann auch die Auswertung der Ergebnisse eine Einschränkung sein. Die qualitative Bewertung der Freitextantworten lässt eine gewisse Subjektivität zu, anders als bei *Multiple Choice*- oder *Long Menu*-Antworten. Die Ergebnisse, die sich auf die inhaltliche Auswertung der Freitextfragen stützen, müssen durch diese Linse betrachtet werden.

Letztendlich ist auch die Übertragbarkeit der Ergebnisse in die klinische Praxis eingeschränkt: obwohl *Key Feature*-Prüfungen *Clinical Reasoning* abbilden können (Farmer und Page 2005; Hrynchak et al. 2014), muss beachtet werden, dass der Transfer vom computerbasierten *Key Feature*-Setting auf echte klinische Situationen eine Hürde bleibt. Beispielsweise bestand bei den *Key Feature*-Prüfungen weniger zeitlicher und emotionaler Stress, es wurde immer nur ein Fall zur selben Zeit bearbeitet, es gab keine Kommunikation mit Mitarbeitern oder Patienten und keine Ablenkungen. Das Übungssetting hat gerade für *Clinical Reasoning*-Anfänger im Medizinstudium seine Berechtigung, ersetzt aber nicht die „hautnahe" klinische Erfahrung.

4.3.2 Stärken

Zu den Stärken dieser Arbeit zählt, dass es sich um eine prospektive randomisierte, kontrollierte Studie handelt. Die Stichprobe umfasste mit 108 Studierenden eine akzeptable Probandenzahl, anhand derer sinnvolle statistische Aussagen getroffen werden können.

Wie bereits in Kapitel 1.6 beschrieben, sind *Key Feature*-Prüfungen ein geeignetes Format zum Prüfen von *Clinical Reasoning*, das Vorteile gegenüber anderen Ansätzen bietet. Durch

den *Testing Effect* sind *Key Feature*-Prüfungen auch für den Einsatz zu Lehrzwecken nutzbar. Die verwendeten *Key Feature*-Fälle entsprachen somit nicht nur einem bereits erwiesenermaßen wirksamen Konzept zum Lernen und Prüfen von *Clinical Reasoning* (Farmer und Page 2005; Hrynchak et al. 2014), sondern die Fälle selber wurden bereits in ihrer Effektivität bestätigt (Raupach et al. 2016) und über mehrere Semester hinweg erprobt und weiterentwickelt. Zudem waren die Inhalte der *Key Feature*-Fälle auf die curriculare Lehre im Studiensemester und auf die Modulklausur 3.1 abgestimmt und konnten somit den Lernverlauf der teilnehmenden Studierenden gut aufgreifen und abbilden.

Für die untersuchte Intervention konnte ein signifikanter Effekt nachgewiesen werden – trotz aller Einschränkungen und Limitationen, die oben erwähnt wurden. Die wichtigste Forschungsfrage 1 konnte somit beantwortet werden. Es hat sich gezeigt, dass die theoretischen Vorteile der bewussten kritischen Auseinandersetzung mit Falschantworten auch in der praktischen Umsetzung nachweisbar sind.

4.4 Ausblick

Die mit dieser Studie belegte Wirksamkeit der Elaboration und Kontrastierung von Falschantworten beim Erwerb von *Clinical Reasoning*-Kompetenz und das Wissen über die Abhängigkeit der Wirksamkeit von bestimmten Faktoren kann für die Zukunft genutzt werden, um die Lehre im klinischen Studienabschnitt des Medizinstudiums um eine weitere Methode zu ergänzen. Die Vorteile dieser Methode liegen darin, dass damit nicht nur Faktenwissen, sondern *Clinical Reasoning* gelernt werden kann, welches bereits im Studium trainiert werden sollte (Norman 2005). Außerdem handelt es sich um eine ressourcensparende Methode, die leicht in bestehende *Key Feature*-Prüfungen – oder möglicherweise auch andere ähnliche Formate – eingebaut werden kann und, nach einmaliger Erstellung, für eine große Zahl Studierender wiederverwendet werden kann. Die erhaltenen Ergebnisse, insbesondere die Freitexte, können dem Lehrenden als Feedback zur Evaluation und Weiterentwicklung der Fragen dienen. Für die Medizinstudierenden bietet dieses Format die Möglichkeit, durch kritisches Hinterfragen und Begründen das Gelernte auf andere Weise zu verfestigen, als es in traditionellen Lehrformen im Medizinstudium der Fall ist. Insbesondere leistungsschwächere Studierende, die im Frontalunterricht vermutlich weniger zu Wort kommen und im Eigenstudium weniger auf eigene Fehler gestoßen werden, scheint dieses Fragenformat im Lernprozess zu unterstützen. Für leistungsstärkere Studierende konnte kein Nachteil nachgewiesen werden. Eine solche Ergänzung der klinischen Lehre dürfte daher einen Gewinn für Lehrende und Lernende darstellen. Es sollte erwogen werden, sie mit ins klinische Curriculum aufzunehmen.

Um die Ergebnisse zu bestätigen und auf eine generalisierbare Ebene zu holen, sollten weitere Studien zu diesem Thema durchgeführt werden mit einer größeren Fallzahl und anderen Studierenden, beispielsweise an anderen Universitäten und mit anderer Semesterzahl, sowie mit *Key Feature*-Fällen aus anderen medizinischen Fachgebieten. Weiter ausgebaut

werden könnte die Intervention durch individuell auf die Antworten der einzelnen Studierenden zugeschnittene Nachfragen, sodass tatsächlich eigene Fehler reflektiert werden. Ohne massiv erhöhten personellen Aufwand könnte dies vielleicht durch den Einsatz künstlicher Intelligenz erzielt werden. Passend dazu könnte ein individualisiertes Feedback eine sinnvolle Ergänzung darstellen.

5 Zusammenfassung

Der Erwerb differentialdiagnostischer und -therapeutischer Kompetenzen stellt ein wesentliches Lernziel im Medizinstudium dar. Unter anderem kann hierzu ein fallbasierter Lernansatz gewählt werden. Aus lernpsychologischer Sicht ist es sinnvoll, klinische Fälle nicht nur diskursiv zu erörtern, sondern die Retention der wesentlichen Aspekte durch testgestützte Verfahren zu fördern. Hierzu bieten sich so genannte *Key Feature*-Fragen an, die Schlüsselstellen im klinischen Management abbilden. Frühere Studien konnten zeigen, dass der studentische Lernerfolg durch den Einsatz solcher Fragen signifikant gesteigert werden kann. Allerdings waren die Lernergebnisse insgesamt noch unbefriedigend, so dass Interventionen zur weiteren Steigerung des Lernerfolgs identifiziert werden müssen. Eine solche Option ist die Aufforderung an Studierende, spezifische Inhalte zu elaborieren und dabei die richtige Antwort von häufigen Falschantworten abzugrenzen. Die Effektivität dieser Maßnahme wurde in dieser Studie untersucht.

An der hier vorgestellten prospektiven, randomisierten Cross-Over-Studie nahmen 108 Medizinstudierende teil, die im Wintersemester 2016/17 im dritten klinischen Semester an der Universitätsmedizin Göttingen studierten. Im Laufe des Semesters nahmen sie parallel zu den Veranstaltungen der Inneren Medizin an 10 computergestützten Fallseminaren teil, in denen sie klinische Fallgeschichten bearbeiteten, in die jeweils *Key Feature*-Fragen eingestreut waren. Auf einige dieser Fragen folgten Freitext-Nachfragen, die auf eine Kontrastierung der richtigen Antwort gegenüber häufigen Falschantworten abzielten („Interventionsitems"). Die insgesamt 30 Interventionsitems wurden gleichmäßig auf die beiden Studierendengruppen aufgeteilt. Die Hälfte der Items wurden in einer Gruppe mitsamt der Freitext-Nachfragen angezeigt; die gleichen Items wurden in der anderen Gruppe ohne Nachfragen angezeigt („Kontrollitems"), und umgekehrt. Am Ende des Semesters und vier Monate später wurden die in den Interventions- und Kontrollitems erzielten Punkte miteinander verglichen.

Am Ende des Semesters bestand ein signifikanter Unterschied in den studentischen Leistungen zwischen Interventions- und Kontrollitems (65,7 ± 19,6% vs. 52,4 ± 22,9%; $p = 0{,}022$; $d = 0{,}16$). Vier Monate später war kein signifikanter Unterschied mehr nachweisbar. In einer univariaten linearen Regression zeigte sich ein signifikanter Zusammenhang zwischen schlechteren Klausurleistungen im Fach Innere Medizin und dem Unterschied zwischen den Leistungen in Interventions- und Kontrollitems: Leistungsschwächere Studierende profitierten stärker von der Intervention. In einer Reihe multivariater linearer Regressionen war diese Assoziation jedoch nicht mehr signifikant.

Die vorliegende Studie zeigt, dass die Ergänzung von *Key Feature*-Prüfungen um ein Element der Elaboration den kurzfristigen studentischen Lernerfolg erhöht, langfristig aber

nicht zu einer Förderung des klinischen Denkens führt. Künftige Studien sollten untersuchen, welche weiteren Maßnahmen die Nachhaltigkeit des Effekts stärken können.

6 Anhang

Anhang 1: Itemschwierigkeiten und -trennschärfe im *Entry Exam*

	Itemschwierigkeit	**Itemtrennschärfe**
Entry Exam		
Fall 1, KF 1: Stellen der Diagnose "stabile Angina pectoris"	0,56 ± 0,498	0,137
Fall 1, KF 2: Ergometrie zum Ischämienachweis	0,53 ± 0,502	0,339
Fall 1, KF 3: sofortige Koronarangiographie bei STEMI	0,38 ± 0,488	0,052
Fall 1, KF 4: Verdachtsdiagnose Perikarditis epistenocardica bei neuer ST-Hebung nach Infarkt	0,09 ± 0,291	0,160
Fall 1, KF 5: Indikation zur primärprophylaktischen ICD-Implantation bei persistierend schlechter LV-Funktion	0,04 ± 0,19	0,264
Fall 2, KF 1: Erkennen einer Obstruktion in der Lungenfunktion	0,17 ± 0,374	-0,246
Fall 2, KF 2: Erkennen einer Lobärpneumonie im Röntgenbild	0,34 ± 0,477	0,204
Fall 2, KF 3: CRB-65-Index zur Entscheidung über die Empfehlung zur stationären Aufnahme	0,01 ± 0,096	-0,005
Fall 2, KF 4: V.a. parapneumonischen Erguss bei typischer Klinik	0,49 ± 0,502	0,220
Fall 2, KF 5: BGA bei V.a. beginnende CO2-Narkose	0,4 ± 0,492	0,081
Fall 2, KF 6: NIV bei nachgewiesener Hyperkapnie	0,15 ± 0,357	0,065
Fall 3, KF 1: Verdacht auf sekundäre Hypertonie bei Manifestation >60 J., diskrepanten Organschäden und Therapierefraktärität	0,04 ± 0,19	0,177
Fall 3, KF 2: Verdacht auf diastolische Dysfunktion bei guter EF, NYHA II und erhöhtem BNP	0,03 ± 0,165	0,205
Fall 3, KF 3: Krankenhaus-Einweisung bei Hyponatriämie <120 mM	0,15 ± 0,357	0,251
Fall 3, KF 4: Erkennen eines Nephrotischen Syndroms anhand der Labor- und Urinbefunde	0,18 ± 0,383	0,290
Fall 3, KF 5: Ableiten der Verdachtsdiagnose "Akute Lungenembolie" bei typischer Klinik und Risikoprofil	0,21 ± 0,411	0,290
Fall 3, KF 6: sofortiges Thorax-CT bei hoher Wahrscheinlichkeit und klinischer Stabilität	0,2 ± 0,405	0,112
Fall 4, KF 1: Schellong-Test bei V.a. orthostatisch bedingte Synkope	0,28 ± 0,45	0,260
Fall 4, KF 2: Erkennen einer Tachyarrhythmia absoluta im EKG	0,28 ± 0,45	0,292
Fall 4, KF 3: Frequenzsenkung bei Tachyarrhythmia absoluta unbekannter Dauer	0,39 ± 0,49	0,251
Fall 4, KF 4: laborchemische Diagnose einer manifesten Hyperthyreose nach Konstrastmittelgabe	0,63 ± 0,485	0,199
Fall 4, KF 5: Absetzen von Amiodaron bei manifester Hyperthyreose	0,32 ± 0,47	0,201
Fall 4, KF 6: klinischer Verdacht auf Lungenfibrose bei Dyspnoe und inspiratorischem Velcro-Knistern	0,2 ± 0,405	0,257
Fall 4, KF 7: Erkennen einer Restriktion in der Lungenfunktionsdiagnostik	0,05 ± 0,211	0,236

Fall 4, KF 8: Indikationsstellung zur LTOT aufgrund eines pO2 <55 mmHg in der arteriellen BGA	0,62 ± 0,488	0,105
Fall 5, KF 1: Verdacht auf supraventrikuläre Genese einer Tachycardie anhand eines beschriebenen EKGs	0,22 ± 0,418	0,198
Fall 5, KF 2: Adenosin zur Unterbrechung einer SVT	0,04 ± 0,19	0,191
Fall 5, KF 3: Elektrophysiologische Untersuchung zur Diagnosesicherung und Therapie	0,14 ± 0,347	0,152
Fall 5, KF 4: Schrittmacher-Implantation bei kompletter AV-Blockierung	0,43 ± 0,497	0,227
Fall 5. KF 5: Verdacht auf Pneumothorax bei entsprechender Klinik nach Punktion der V. subclavia	0,62 ± 0,488	0,341

Anhang 2: Itemschwierigkeit und -trennschärfe im *Exit Exam*

	Itemschwierigkeit	**Itemtrennschärfe**
Exit Exam		
Fall 1, KF 1: Stellen der Diagnose "stabile Angina pectoris"	0,84 ± 0,366	0,374
Fall 1, KF 2: Ergometrie zum Ischämienachweis	0,71 ± 0,454	0,441
Fall 1, KF 3: sofortige Koronarangiographie bei STEMI	0,81 ± 0,39	0,274
Fall 1, KF 4: Verdachtsdiagnose Perikarditis epistenocardica bei neuer ST-Hebung nach stattgehabtem Infarkt	0,65 ± 0,48	0,446
Fall 1, KF 5: Indikation zur primärprophylaktischen ICD-Implantation bei persistierend schlechter LV-Funktion	0,6 ± 0,492	0,239
Fall 2, KF 1: Erkennen einer Obstruktion in der Lungenfunktion	0,61 ± 0,49	0,213
Fall 2, KF 2: Erkennen einer Lobärpneumonie im Röntgenbild	0,6 ± 0,492	0,400
Fall 2, KF 3: CRB-65-Index zur Entscheidung über die Empfehlung zur stationären Aufnahme	0,78 ± 0,418	0,368
Fall 2, KF 4: V.a. parapneumonischen Erguss bei typischer Klinik	0,85 ± 0,357	0,409
Fall 2, KF 5: BGA bei V.a. beginnende CO2-Narkose	0,75 ± 0,435	0,319
Fall 2, KF 6: NIV bei nachgewiesener Hyperkapnie	0,6 ± 0,492	0,242
Fall 3, KF 1: Verdacht auf sekundäre Hypertonie bei Manifestation >60 J., diskrepanten Organschäden und Therapierefraktärität	0,39 ± 0,49	0,320
Fall 3, KF 2: Verdacht auf diastolische Dysfunktion bei guter EF, NYHA II und erhöhtem BNP	0,19 ± 0,39	0,330
Fall 3, KF 3: Krankenhaus-Einweisung bei Hyponatriämie <120 mM	0,56 ± 0,498	0,419
Fall 3, KF 4: Erkennen eines Nephrotischen Syndroms anhand der Labor- und Urinbefunde	0,77 ± 0,424	0,416
Fall 3, KF 5: Ableiten der Verdachtsdiagnose "Akute Lungenembolie" bei typischer Klinik und Risikoprofil	0,58 ± 0,495	0,459
Fall 3, KF 6: sofortiges Thorax-CT bei hoher Wahrscheinlichkeit und klinischer Stabilität	0,44 ± 0,499	0,238
Fall 4, KF 1: Schellong-Test bei V.a. orthostatisch bedingte Synkope	0,64 ± 0,483	0,294
Fall 4, KF 2: Erkennen einer Tachyarrhythmia absoluta im EKG	0,59 ± 0,494	0,464
Fall 4, KF 3: Frequenzsenkung bei Tachyarrhythmia absoluta unbekannter Dauer	0,56 ± 0,499	0,504
Fall 4, KF 4: laborchemische Diagnose einer manifesten Hyperthyreose nach Konstrastmittelgabe	0,83 ± 0,374	0,377

Fall 4, KF 5: Absetzen von Amiodaron bei manifester Hyperthyreose	0,8 ± 0,405	0,519
Fall 4, KF 6: klinischer Verdacht auf Lungenfibrose bei Dyspnoe und inspiratorischem Velcro-Knistern	0,67 ± 0,474	0,488
Fall 4, KF 7: Erkennen einer Restriktion in der Lungenfunktionsdiagnostik	0,21 ± 0,411	0,158
Fall 4, KF 8: Indikationsstellung zur LTOT aufgrund eines pO2 <55 mmHg in der arteriellen BGA	0,54 ± 0,501	0,259
Fall 5, KF 1: Verdacht auf supraventrikuläre Genese einer Tachycardie anhand eines beschriebenen EKGs	0,6 ± 0,492	0,438
Fall 5, KF 2: Adenosin zur Unterbrechung einer SVT	0,65 ± 0,48	0,549
Fall 5, KF 3: Elektrophysiologische Untersuchung zur Diagnosesicherung und Therapie	0,8 ± 0,405	0,556
Fall 5, KF 4: Schrittmacher-Implantation bei kompletter AV-Blockierung	0,75 ± 0,435	0,426
Fall 5. KF 5: Verdacht auf Pneumothorax bei entsprechender Klinik nach Punktion der V. subclavia	0,83 ± 0,374	0,478

Anhang 3: Itemschwierigkeit und -trennschärfe im *Retention Test*

	Itemschwierigkeit	**Itemtrennschärfe**
Retention Test		
Fall 1, KF 1: Stellen der Diagnose "stabile Angina pectoris"	0,83 ± 0,374	0,330
Fall 1, KF 2: Ergometrie zum Ischämienachweis	0,77 ± 0,424	0,447
Fall 1, KF 3: sofortige Koronarangiographie bei STEMI	0,78 ± 0,418	0,338
Fall 1, KF 4: Verdachtsdiagnose Perikarditis epistenocardica bei neuer ST-Hebung nach stattgehabtem Infarkt	0,55 ± 0,5	0,392
Fall 1, KF 5: Indikation zur primärprophylaktischen ICD-Implantation bei persistierend schlechter LV-Funktion	0,34 ± 0,477	0,34
Fall 2, KF 1: Erkennen einer Obstruktion in der Lungenfunktion	0,48 ± 0,502	0,538
Fall 2, KF 2: Erkennen einer Lobärpneumonie im Röntgenbild	0,61 ± 0,49	0,389
Fall 2, KF 3: CRB-65-Index zur Entscheidung über die Empfehlung zur stationären Aufnahme	0,62 ± 0,488	0,270
Fall 2, KF 4: V.a. parapneumonischen Erguss bei typischer Klinik	0,73 ± 0,445	0,509
Fall 2, KF 5: BGA bei V.a. beginnende CO2-Narkose	0,76 ± 0,43	0,412
Fall 2, KF 6: NIV bei nachgewiesener Hyperkapnie	0,46 ± 0,501	0,304
Fall 3, KF 1: Verdacht auf sekundäre Hypertonie bei Manifestation >60 J., diskrepanten Organschäden und Therapierefraktärität	0,46 ± 0,501	0,323
Fall 3, KF 2: Verdacht auf diastolische Dysfunktion bei guter EF, NYHA II und erhöhtem BNP	0,19 ± 0,39	0,361
Fall 3, KF 3: Krankenhaus-Einweisung bei Hyponatriämie <120 mM	0,49 ± 0,502	0,414
Fall 3, KF 4: Erkennen eines Nephrotischen Syndroms anhand der Labor- und Urinbefunde	0,8 ± 0,405	0,474
Fall 3, KF 5: Ableiten der Verdachtsdiagnose "Akute Lungenembolie" bei typischer Klinik und Risikoprofil	0,56 ± 0,499	0,466
Fall 3, KF 6: sofortiges Thorax-CT bei hoher Wahrscheinlichkeit und klinischer Stabilität	0,38 ± 0,488	0,258

Fall 4, KF 1: Schellong-Test bei V.a. orthostatisch bedingte Synkope	0,55 ± 0,5	0,231
Fall 4, KF 2: Erkennen einer Tachyarrhythmia absoluta im EKG	0,54 ± 0,501	0,444
Fall 4, KF 3: Frequenzsenkung bei Tachyarrhythmia absoluta unbekannter Dauer	0,52 ± 0,502	0,357
Fall 4, KF 4: laborchemische Diagnose einer manifesten Hyperthyreose nach Konstrastmittelgabe	0,84 ± 0,366	0,367
Fall 4, KF 5: Absetzen von Amiodaron bei manifester Hyperthyreose	0,85 ± 0,357	0,462
Fall 4, KF 6: klinischer Verdacht auf Lungenfibrose bei Dyspnoe und inspiratorischem Velcro-Knistern	0,69 ± 0,463	0,445
Fall 4, KF 7: Erkennen einer Restriktion in der Lungenfunktionsdiagnostik	0,09 ± 0,291	0,179
Fall 4, KF 8: Indikationsstellung zur LTOT aufgrund eines pO2 <55 mmHg in der arteriellen BGA	0,55 ± 0,5	0,244
Fall 5, KF 1: Verdacht auf supraventrikuläre Genese einer Tachycardie anhand eines beschriebenen EKGs	0,6 ± 0,492	0,296
Fall 5, KF 2: Adenosin zur Unterbrechung einer SVT	0,61 ± 0,49	0,574
Fall 5, KF 3: Elektrophysiologische Untersuchung zur Diagnosesicherung und Therapie	0,76 ± 0,43	0,501
Fall 5, KF 4: Schrittmacher-Implantation bei kompletter AV-Blockierung	0,82 ± 0,383	0,410
Fall 5. KF 5: Verdacht auf Pneumothorax bei entsprechender Klinik nach Punktion der V. subclavia	0,82 ± 0,383	0,427

Anhang 4: Häufigkeiten ausgewählter Falschantworten in *Entry Exam*, *Exit Exam* und *Retention Test*. Farbig hinterlegt sind die in der Freitextfrage erwähnten Falschantworten, wenn vorhanden.

Item	**Häufigkeit in Prozent** (n=108)						
	Antwort	***Entry Exam***		***Exit Exam***		***Retention Test***	
		Intervention	Kontrolle	Intervention	Kontrolle	Intervention	Kontrolle
Fall 1 KF 1: Diagnose stabile Angina pectoris (Interventionsitem Gruppe A)	Richtige Antwort	52,8%	60,0 %	88,7%	80,0 %	83,0%	83,6 %
	2 Instabile Angina pectoris	32,1%	25,5 %	5,7%	16,4 %	3,8%	12,7 %
	3 Belastungsangina	15,1%	12,7 %	0,0%	0,0%	3,8%	0,0%
	4 KHK/Akuter Myokardinfarkt	0,0%	1,8%	3,8%	1,8%	0,0%	1,8%
	Andere/keine Antwort	0,0%	0,0%	1,9%	1,8%	9,4%	1,8%
Fall 1 KF2: Ergometrie zum Ischämienachweis (Interventionsitem Gruppe A)	Richtige Antwort	52,8%	52,7 %	73,6%	69,1 %	73,6%	80,0 %
	2 Echokardiographie	7,5%	14,5 %	15,1%	16,4 %	3,8%	10,9 %
	3 EKG	20,8%	12,7 %	5,7%	5,5%	9,4%	3,6%

	4 Laborwert-Kontrolle	13,2%	9,1%	0,0%	0,0%	1,9%	1,8%
	5 Koronarangiographie	5,7%	3,6%	3,8%	3,6%	5,7%	0,0%
	6	0,0%	3,6%	0,0%	0,0%		
	Andere/keine Antwort	0,0%	3,6%	1,9%	5,5%	5,7%	3,6%
Fall 1 KF3: Coro-Indikation bei STEMI (Interventionsitem Gruppe A)	Richtige Antwort	37,7%	38,2 %	86,8%	76,4 %	71,7%	83,6 %
	2 Laborwertbestimmung (Troponine)	50,9%	43,6 %	5,7%	10,9 %	3,8%	5,5%
	3 Echokardiographie	3,8%	5,5%	5,7%	9,1%	11,3%	7,3%
	4 EKG	7,5%	12,7 %	1,9%	1,8%	5,7%	1,8%
	Andere/keine Antwort	0,0%	0,0%	0,0%	1,8%	7,5%	1,8%
Fall 1 KF4: Pericarditis epistenocardica (Interventionsitem Gruppe B)	Richtige Antwort	9,1%	9,4%	69,1%	60,4 %	60,0%	49,1 %
	2 Myokardinfarkt	76,4%	69,8 %	5,5%	15,1 %	12,7%	18,9 %
	3 Stentverschluss/Re-Infarkt	10,9%	13,2 %	1,8%	5,7%	1,8%	1,9%
	4 Myokarditis	0,0%	3,8%	10,9%	5,7%	5,5%	13,2 %
	5 Perikardtamponade	0,0%	0,0%	7,3%	0,0%	0,0%	0,0%
	Andere/keine Antwort	3,6%	3,8%	5,5%	13,2 %	20,0%	17,0 %
Fall 1 KF5: Indikation zur ICD-Implantation (Interventionsitem Gruppe A)	Richtige Antwort	1,9%	5,5%	64,2%	56,4 %	30,2%	38,2 %
	2 Schrittmacher-Implantation	0,0%	3,6%	17,0%	18,2 %	39,6%	27,3 %
	3 medikamentöse Therapie	35,8%	25,5 %	5,7%	7,3%	7,5%	9,1%
	4 weitere Diagnostik	7,5%	10,9 %	0,0%	0,0%	0,0%	0,0%
	5 abwarten	3,8%	3,6%	0,0%	1,8%	0,0%	0,0%
	6 Sport	15,1%	16,4 %	3,8%	0,0%	5,7%	3,6%
	7	0,0%	0,0%	3,8%	1,8%	0,0%	0,0%
	Andere/keine Antwort	35,8%	34,5 %	5,7%	14,5 %	17,0%	21,8 %
Fall 2 KF1: Erkennen einer Obstruktion in der Lungenfunktion (Interventionsitem Gruppe B)	Richtige Antwort	10,9%	22,6 %	67,3%	54,7 %	58,2%	37,7 %
	2 Einsekundenkapazität (% des Solls)	38,2%	30,2 %	9,1%	15,1 %	14,5%	22,6 %
	3 Einsekundenkapazität/Vitalkapazität (% des	32,7%	18,9 %	21,8%	13,2 %	16,4%	9,4%

	Solls)						
	4 Einsekundenkapazität (Ist-Wert)	3,6%	9,4%	1,8%	5,7%	7,3%	18,9 %
	5 Totale Lungenkapazität (% des Solls)	3,6%	1,9%	0,0%	3,8%	0,0%	1,9%
	Andere/keine Antwort	10,9%	17,0 %	0,0%	7,5%	3,6%	9,4%
Fall 2 KF2: Erkennen einer Lobärpneumonie im Röntgenbild (Interventionsitem Gruppe B)	Richtige Antwort	34,5%	34,0 %	65,5%	54,7 %	61,8%	60,4 %
	2 Pneumonie, allgemein	47,3%	39,6 %	18,2%	24,5 %	23,6%	26,4 %
	3 Pneumonie, spezifisch	5,5%	15,1 %	7,3%	7,5%	9,1%	3,8%
	4 Emphysem	3,6%	0,0%	1,8%	0,0%	0,0%	0,0%
	5 Lungenödem	0,0%	3,8%	1,8%	0,0%	0,0%	3,8%
	Andere/keine Antwort	9,1%	7,5%	5,5%	13,2 %	5,5%	5,7%
Fall 2 KF3: CRB-65-Index (Interventionsitem Gruppe A)	Richtige Antwort	1,9%	0,0%	79,2%	76,4 %	56,6%	67,3 %
	2 Laborwerte (diverse)	67,9%	61,8 %	3,8%	1,8%	7,5%	3,6%
	3 O2-Sättigung	11,3%	14,5 %	0,0%	0,0%	1,9%	0,0%
	4 Fieber	3,8%	9,1%	0,0%	0,0%	1,9%	3,6%
	5 CHA2D2-VASc-Score	0,0%	0,0%	1,9%	1,8%	9,4%	5,5%
	Andere/keine Antwort	15,1%	14,5 %	15,1%	20,0 %	22,6%	20,0 %
Fall 2 KF4: V.a. parapneumonischen Erguss (Interventionsitem Gruppe A)	Richtige Antwort	47,2%	50,9 %	84,9%	85,5 %	73,6%	72,7 %
	2 Lungenödem	17,0%	18,2 %	5,7%	7,3%	7,5%	7,3%
	3 Pneumonie	9,4%	12,7 %	0,0%	0,0%	1,9%	9,1%
	4 Pneumothorax	1,9%	1,8%	1,9%	3,6%	1,9%	7,3%
	5 Emphysem	3,8%	7,3%	0,0%	0,0%	0,0%	0,0%
	Andere/keine Antwort	20,8%	9,1%	7,5%	3,6%	15,1%	3,6%
Fall 2 KF5: BGA bei V.a. beginnende CO2-Narkose (Interventionsitem Gruppe B)	Richtige Antwort	43,6%	35,8 %	80,0%	69,8 %	80,0%	71,7 %
	2 Röntgen-Thorax	3,6%	1,9%	1,8%	0,0%	1,8%	3,8%
	3 Laborwertkontrolle	7,3%	13,2 %	9,1%	11,3 %	12,7%	7,5%
	4 Sonstige Bildgebung (radiologisch)	14,5%	26,4 %	1,8%	1,9%	1,8%	3,8%

	Andere/keine Antwort	30,9%	22,6 %	7,3%	17,0 %	3,6%	13,2 %
Fall 2 KF6: NIV bei Hyperkapnie (Interventionsitem Gruppe A)	Richtige Antwort	17,0%	12,7 %	54,7%	65,5 %	39,6%	52,7 %
	2 Sauerstoff-Gabe	22,6%	21,8 %	9,4%	16,4 %	13,2%	12,7 %
	3 CPAP	13,2%	14,5 %	15,1%	7,3%	20,8%	21,8 %
	4 Pufferung/Azidoseausgleich	11,3%	10,9 %	1,9%	5,5%	5,7%	9,1%
	5 Intubation	15,1%	14,5 %	7,5%	1,8%	5,7%	0,0%
	6	0,0%	0,0%			3,8%	1,8%
	Andere/keine Antwort	20,8%	25,5 %	11,3%	3,6%	11,3%	1,8%
Fall 3 KF1: V.a. sekundäre Hypertonie (Interventionsitem Gruppe B)	Richtige Antwort	1,8%	5,7%	40,0%	37,7 %	49,1%	43,4 %
	2 Sonstige/allgemeine Hypertonie	10,9%	18,9 %	10,9%	22,6 %	16,4%	11,3 %
	3 Diabetes mellitus	34,5%	20,8 %	12,7%	5,7%	12,7%	13,2 %
	4 Nierenkrankheit	5,5%	7,5%	0,0%	13,2 %		
	5 Glaukom	16,4%	15,1 %	5,5%	5,7%	5,5%	5,7%
	Andere/keine Antwort	30,9%	32,1 %	30,9%	26,4 %	16,4%	26,4 %
Fall 3 KF2: V.a. diastolische Dysfunktion (Interventionsitem Gruppe A)	Richtige Antwort	5,7%	0,0%	20,8%	16,4 %	17,0%	20,0 %
	2 Herzinsuffizienz (Allgemein oder Linksherz-)	26,4%	23,6 %	30,2%	43,6 %	39,6%	34,5 %
	3 Rechtsherzinsuffizienz	24,5%	34,5 %	18,9%	9,1%	18,9%	25,5 %
	4 Kardiomyopathie	1,9%	3,6%	13,2%	7,3%	3,8%	3,6%
	5 Niereninsuffizienz	5,7%	12,7 %	3,8%	5,5%	1,9%	3,6%
	Andere/keine Antwort	35,8%	25,5 %	13,2%	18,2 %	18,9%	12,7 %
Fall 3 KF3: Krankenhaus-Einweisung bei Hyponatriämie (Interventionsitem Gruppe B)	Richtige Antwort	12,7%	17,0 %	61,8%	50,9 %	**60,0%**	**37,7 %**
	2 Natrium-Substitution	23,6%	26,4 %	1,8%	15,1 %	**9,1%**	**15,1 %**
	3 Diuretikum absetzen	14,5%	24,5 %	25,5%	26,4 %	**25,5%**	**34,0 %**
	4 Diuretikum ansetzen	21,8%	5,7%	1,8%	0,0%	**0,0%**	**0,0%**
	5 ACE-Hemmer absetzen	5,5%	7,5%	0,0%	1,9%	**3,6%**	**0,0%**
	Andere/keine Antwort	21,8%	18,9	9,1%	5,7%	**1,8%**	**13,2**

			%				**%**
Fall 3 KF4: Nephrotisches Syndrom (Interventionsitem Gruppe A)	Richtige Antwort	20,8%	14,5 %	81,1%	72,7 %	77,4%	81,8 %
	2 Nephritisches Syndrom	9,4%	5,5%	7,5%	7,3%	9,4%	1,8%
	3 Glomerulonephritis	3,8%	3,6%	0,0%	9,1%	0,0%	1,8%
	4 Niereninsuffizienz/Nierenkrankheit allgemein	20,8%	23,6 %	5,7%	5,5%	5,7%	5,5%
	5 Proteinurie	34,0%	30,9 %	1,9%	1,8%	3,8%	5,5%
	Andere/keine Antwort	11,3%	21,8 %	3,8%	3,6%	3,8%	3,6%
Fall 3 KF5: Ableiten der Verdachtsdiagnose Akute Lungenembolie (Interventionsitem Gruppe B)	Richtige Antwort	12,7%	30,2 %	56,4%	60,4 %	58,2%	52,8 %
	2 Herzinsiffuzienz	41,8%	30,2 %	5,5%	1,9%	5,5%	5,7%
	3 Akutes Koronarsyndrom	5,5%	3,8%	0,0%	1,9%	0,0%	1,9%
	4 Thrombose	9,1%	13,2 %	30,9%	20,8 %	25,5%	22,6 %
	5 Niereninsuffizienz	7,3%	7,5%	1,8%	7,5%	1,8%	3,8%
	Andere/keine Antwort	23,6%	15,1 %	5,5%	7,5%	9,1%	13,2 %
Fall 3 KF6: Thorax-CT bei Lungenembolie-Fall (Interventionsitem Gruppe A)	Richtige Antwort	17,0%	23,6 %	43,4%	45,5 %	**32,1%**	**43,6 %**
	2 Lyse	18,9%	14,5 %	3,8%	7,3%	**9,4%**	**3,6%**
	3 Antikoagulation	18,9%	14,5 %	17,0%	18,2 %	**1,9%**	**25,5 %**
	4 Laborwertbestimmung	5,7%	7,3%	1,9%	1,8%	**0,0%**	**1,8%**
	5 Echokardiographie	3,8%	0,0%	7,5%	5,5%	**9,4%**	**0,0%**
	6 Röntgen	0,0%	3,6%	0,0%	5,5%		
	Andere/keine Antwort	35,8%	36,4 %	26,4%	16,4 %	**47,2%**	**25,5 %**
Fall 4 KF1: Schellong-Test bei V.a. orthostatisch bedingte Synkope (Interventionsitem Gruppe B)	Richtige Antwort	24,5%	30,9 %	65,5%	62,3 %	60,0%	49,1 %
	2 Kipptisch-Untersuchung	24,5%	32,7 %	23,6%	26,4 %	29,1%	43,4 %
	3 Langzeit-Blutdruckmessung	28,3%	14,5 %	0,0%	3,8%	1,8%	3,8%
	4 Echokardiographie	3,8%	1,8%	0,0%	0,0%	0,0%	0,0%

	Andere/keine Antwort	20,0%	18,9 %	10,9%	7,5%	9,1%	3,8%
Fall 4 KF2: Erkennen einer Tachyarrhythmia absoluta im EKG (Interventionsitem Gruppe A)	Richtige Antwort	30,2%	25,5 %	**69,8%**	**49,1 %**	45,3%	61,8 %
	2 AV-Block	7,5%	21,8 %	**9,4%**	**25,5 %**	15,1%	16,4 %
	3 Arrhythmie/Tachykardie allgemein	35,8%	32,7 %	**5,7%**	**16,4 %**	15,1%	12,7 %
	Andere/keine Antwort	26,4%	20,0 %	**15,1%**	**9,1%**	24,5%	9,1%
Fall 4 KF3: Frequenzsenkung bei TAA (Interventionsitem Gruppe B)	Richtige Antwort	36,4%	41,5 %	60,0%	50,9 %	47,3%	56,6 %
	2 Rhythmuskontrolle/Rhythmisierung	23,6%	26,4 %	20,0%	34,0 %	36,4%	24,5 %
	3 Antikoagulation	5,5%	11,3 %	5,5%	0,0%	3,6%	1,9%
	4 Schrittmacher	3,6%	1,9%	0,0%	3,8%	1,8%	3,8%
	Andere/keine Antwort	30,9%		14,5%	11,3 %	10,9%	13,2 %
Fall 4 KF4: laborchemische Diagnose einer manifesten Hyperthyreose (Interventionsitem Gruppe B)	Richtige Antwort	70,9%	54,7 %	89,1%	77,4 %	89,1%	79,2 %
	2 Hypothyreose	9,1%	15,1 %	3,6%	13,2 %	3,6%	3,8%
	3 latente Hyperthyreose	1,8%	3,8%	3,6%	1,9%	1,8%	1,9%
	Andere/keine Antwort	18,2%	26,4 %	3,6%	7,5%	5,5%	15,1 %
Fall 4 KF5: Absetzen von Amiodaron (Interventionsitem Gruppe B)	Richtige Antwort	32,7%	32,1 %	83,6%	75,5 %	**94,5%**	**75,5 %**
	2 Thyreoidektomie	7,3%	15,1 %	1,8%	5,7%	**3,6%**	**3,8%**
	3 Thyreostatika	1,8%	9,4%	1,8%	0,0%	**0,0%**	**1,9%**
	4 Radionuklidtherapie	1,8%	3,8%	0,0%	1,9%		
	5	1,8%	0,0%	0,0%	0,0%		
	Andere/keine Antwort	54,5%		12,7%	17,0 %	**1,8%**	**18,9 %**
Fall 4 KF6: Verdacht auf Lungenfibrose (Interventionsitem Gruppe A)	Richtige Antwort	17,0%	23,6 %	60,4%	72,7 %	56,6%	81,8 %
	2 Emphysem	18,9%	23,6 %	7,5%	5,5%	7,5%	1,8%
	3 Lungenödem	28,3%	12,7 %	13,2%	7,3%	7,5%	5,5%

	4 Pneumonie	13,2%	16,4 %	3,8%	5,5%	9,4%	5,5%
	5 Lungenarterienembolie	7,5%	3,6%	0,0%	0,0%	0,0%	0,0%
	Andere/keine Antwort	15,1%	20,0 %	15,1%	9,1%	18,9%	5,5%
Fall 4 KF7: TLC zur Diagnostik einer pulmonalen Restriktion (Interventionsitem Gruppe B)	Richtige Antwort	1,8%	7,5%	**7,3%**	**35,8 %**	10,9%	7,5%
	2 Vitalkapazität (% des Solls und Ist-Wert)	52,7%	37,7 %	**45,5%**	**20,8 %**	34,5%	41,5 %
	3 FEV1/Vitalkapazität (% des Solls und Istwert)	25,5%	22,6 %	**20,0%**	**18,9 %**	21,8%	22,6 %
	5 FEV (% des Solls und Ist-Wert)	10,9%	13,2 %	**3,6%**	**7,5%**	12,7%	11,3 %
	6 Residualvolumen/Totale Lungenkapazität (% des Solls und Ist-Wert)	1,8%	9,4%	**1,8%**	**3,8%**	7,3%	1,9%
	7 Totale Lungenkapazität (Ist-Wert)	5,5%	0,0%	**18,2%**	**9,4%**	10,9%	7,5%
	Andere/keine Antwort	1,8%	9,4%	**3,6%**	**3,8%**	1,8%	7,5%
Fall 4 KF8: Indikation zur LTOT (Interventionsitem Gruppe A)	Richtige Antwort	62,3%	61,8 %	**50,9%**	**56,4 %**	50,9%	58,2 %
	2 (Nicht-invasive) Beatmung	13,2%	9,1%	**26,4%**	**20,0 %**	18,9%	20,0 %
	3 CPAP	3,8%	1,8%	**11,3%**	**0,0%**	13,2%	5,5%
	Andere/keine Antwort	20,8%	27,3 %	**11,3%**	**23,6 %**	17,0%	16,4 %
Fall 5 KF1: Erkennen einer supraventrikulären Tachykardie (Interventionsitem Gruppe B)	Richtige Antwort	21,8%	22,6 %	60,0%	60,4 %	61,8%	58,5 %
	2 TAA bei Vorhofflimmern	23,6%	26,4 %	14,5%	18,9 %	20,0%	15,1 %
	3 Ventrikuläre Tachykardie	9,1%	13,2 %	14,5%	7,5%	12,7%	9,4%
	4 Tachykardie	23,6%	15,1 %	3,6%	3,8%	1,8%	0,0%
	5 AV-Block	10,9%	5,7%	0,0%	0,0%	0,0%	3,8%
	Andere/keine Antwort	10,9%	17,0 %	7,3%	9,4%	3,6%	13,2 %
Fall 5 KF2: Adenosin zur Terminierung einer Tachykardie (Interventionsitem Gruppe A)	Richtige Antwort	5,7%	1,8%	58,5%	70,9 %	52,8%	69,1 %
	2 Betablocker	13,2%	16,4 %	0,0%	1,8%	0,0%	3,6%
	3 Amiodaron	1,9%	3,6%	7,5%	5,5%	5,7%	3,6%

	4 Kardioversion	15,1%	16,4 %	7,5%	1,8%	1,9%	3,6%
	5 medikamentöse Therapie (sonstige)	28,3%	27,3 %	11,3%	7,3%	24,5%	12,7 %
	Andere/keine Antwort	35,8%	34,5 %	15,1%	12,7 %	15,1%	7,3%
Fall 5 KF3: EPU zur Diagnose und Therapie (Interventionsitem Gruppe B)	Richtige Antwort	14,5%	13,2 %	89,1%	69,8 %	81,8%	69,8 %
	2 Langzeit-EKG	40,0%	24,5 %	1,8%	3,8%	5,5%	1,9%
	3 Echokardiographie	14,5%	26,4 %	0,0%	7,5%	3,6%	3,8%
	4 Belastungs-EKG	5,5%	5,7%	0,0%	0,0%	0,0%	1,9%
	5 Herzkatheter/Koronarangiographie	1,8%	5,7%	0,0%	0,0%	1,8%	5,7%
	6 Elektrokardioversion	0,0%	0,0%	1,8%	5,7%	3,6%	1,9%
	Andere/keine Antwort	23,6%	24,5 %	7,3%	13,2 %	3,6%	15,1 %
Fall 5 KF4: Schrittmacher-Implantation bei kompletter AV-Blockierung (Interventionsitem Gruppe A)	Richtige Antwort	43,4%	41,8 %	69,8%	80,0 %	79,2%	85,5 %
	2 Elektrokardioversion	15,1%	12,7 %	5,7%	1,8%	5,7%	1,8%
	3 ICD-Implantation	3,8%	3,6%	9,4%	7,3%	1,9%	3,6%
	4 Medikamentöse Therapie (diverse)	28,3%	21,8 %	1,9%	1,8%	3,8%	1,8%
	5 Passagerer Schrittmacher	1,9%	1,8%	3,8%	5,5%	0,0%	0,0%
	6 Biventrikulärer Schrittmacher	0,0%	1,8%	3,8%	1,8%	1,9%	1,8%
	Andere/keine Antwort	7,5%	16,4 %	5,7%	1,8%	7,5%	5,5%
Fall 5 KF5: V.a. Pneumothorax (Interventionsitem Gruppe B)	Richtige Antwort	74,5%	49,1 %	90,9%	75,5 %	89,1%	75,5 %
	2 Lungenembolie	3,6%	5,7%	0,0%	3,8%	0,0%	3,8%
	3 Lungenödem	5,5%	9,4%	1,8%	5,7%	0,0%	0,0%
	4 Pleuraerguss	1,8%	1,9%	3,6%	1,9%	7,3%	3,8%
	5 Pneumonie	1,8%	3,8%	0,0%	1,9%	1,8%	5,7%
	Andere/keine Antwort	12,7%	30,2 %	3,6%	11,3 %	1,8%	11,3 %

Anhang 5: Interventionsitems mit Wortlaut der Freitextfragen und durchschnittlicher Freitextlänge und Freitextbewertung.

Item	**Freitextfrage**	**Mittlere Freitextlänge (Zeichen)**		**Mittlere Freitextbewertung (0-3 Punkte)**	
		1. Präsentation	2. Präsentation	1. Präsentation	2. Präsentation
Stellen der Diagnose „stabile Angina pectoris“	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Zur Erhöhung Ihres Lernerfolgs erläutern Sie bitte kurz, durch welche Kriterien eine stabile von einer instabilen Angina pectoris unterschieden werden kann.	77,33	71,12	0,75	0,82
Ergometrie zum Ischämienachweis	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Zur Erhöhung Ihres Lernerfolgs erklären Sie bitte kurz den Unterschied zwischen einer transthorakalen Echokardiographie in Ruhe und einer Ergometrie hinsichtlich der Aussagekraft über eine Koronarinsuffizienz.	90,08	85,80	0,6	0,65
sofortige Koronarangiographie bei STEMI	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Warum sollte nun umgehend eine Herzkatheteruntersuchung erfolgen, ohne zuvor die Laborergebnisse abzuwarten oder eine Echokardiographie durchzuführen?	64,71	59,25	0,69	0,58
Verdachtsdiagnose Perikarditis epistenocardica bei neuer ST-Hebung nach stattgehabtem Infarkt	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Welche Befunde sprechen in diesem Fall eher für eine Perikarditis epistenocardica und gegen das Vorliegen eines Myokardinfark-	38,39	41,76	0,44	0,52

	tes oder eines Herzwandaneurysmas?				
Erkennen einer Obstruktion in der Lungenfunktion	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Bitte erklären Sie in Stichworten, warum der Quotient FEV1/VC (Ist-Wert) bei der Frage nach dem Vorliegen einer bronchialen Obstruktion aussagekräftiger ist als der Quotient FEV1/VC (% vom Soll) oder die FEV1 (Ist-Wert oder % vom Soll).	35,24	33,54	0,15	0,27
Erkennen einer Lobärpneumonie im Röntgenbild	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet; insbesondere wurde oft nicht die richtige Lokalisation der Pneumonie angegeben. Bitte erläutern Sie kurz, warum in diesem Fall eine Pneumonie des linken Oberlappens und keine Mittellappenpneumonie vorliegt.	26,57	42,94	0,69	0,58
CRB-65-Index zur Entscheidung über die Empfehlung zur stationären Aufnahme	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Bitte tragen Sie in folgendes Freitextfeld die 4 Kriterien ein, die in den CRB-65-Score eingehen.	35,55	24,41	0,82	0,56
V.a. parapneumonischen Erguß bei typischer Klinik	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Um Ihren Lernerfolg zu erhöhen, skizzieren Sie bitte kurz, was in diesem Fall für das Vorliegen eines Pleuraergusses spricht.	29,48	37,54	0,53	0,63

BGA bei V.a. beginnende CO2-Narkose	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Bitte erläutern Sie kurz, warum in dieser Situation primär eine Blutgasanalyse und kein Röntgenbild angefertigt werden sollte.	48,52	44,36	0,58	0,53
NIV bei nachgewiesener Hyperkapnie	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Warum sollte in dieser Situation auf keinen Fall eine reine Sauerstoffgabe erfolgen?	44,54	42,21	0,66	0,5
Verdacht auf sekundäre Hypertonie bei Manifestation >60 J., diskrepanten Organschäden und Therapierefraktärität	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Zur Erhöhung Ihres Lernerfolgs listen Sie bitte in Stichworten auf, welche anamnestischen Hinweise und Befunde darauf hinweisen, dass bei Frau Zeiher eine sekundäre Hypertonie (und keine Kardiomyopathie oder Herzinsuffizienz) vorliegt?	42,31	33,69	0,85	0,36
Verdacht auf diastolische Dysfunktion bei guter EF, NYHA II und erhöhtem BNP	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Zur Steigerung Ihres Lernerfolgs geben Sie bitte in folgendes Freitextfeld in Stichworten ein, woran Sie in dem geschilderten Fall erkennen können, dass eine diastolische und keine systolische Herzinsuffizienz vorliegt?	34,21	39,92	0,58	0,74
Krankenhaus-Einweisung bei Hyponatriämie <120 mM	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet.	35,93	51,10	0,7	0,78

	Um Ihren Lernerfolg zu erhöhen, skizzieren Sie bitte kurz, warum in dieser Situation zunächst keine Natrium-Substitution erfolgen sollte				
Erkennen eines Nephrotischen Syndroms anhand der Labor- und Urinbefunde	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Um Ihren Lernerfolg zu erhöhen, erläutern Sie bitte kurz, wie sich ein Nephrotisches Syndrom von einem Nephritischen Syndrom unterscheidet.	66,00	58,32	0,64	0,69
Ableiten der Verdachtsdiagnose "Akute Lungenembolie" bei typischer Klinik und Risikoprofil	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Um Ihren Lernerfolg zu erhöhen, skizzieren Sie bitte kurz, warum in diesem Fall am ehesten eine akute Lungenarterienembolie und keine Aortendissektion oder ein Myokardinfarkt vorliegt.	34,71	48,74	0,43	0,45
sofortiges Thorax-CT bei hoher Wahrscheinlichkeit und klinischer Stabilität	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Um Ihren Lernerfolg zu erhöhen, skizzieren Sie bitte kurz, warum ist in dieser Situation keine Lysetherapie indiziert ist.	26,64	30,18	0,31	0,37
Schellong-Test bei V.a. orthostatisch bedingte Synkope	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Bitte erläutern Sie kurz, welche Dysregulation mit dem Schellong-Test nachgewiesen werdenkann und welche andere Dysregulation mit Hilfe einer	53,93	29,21	0,47	0,48

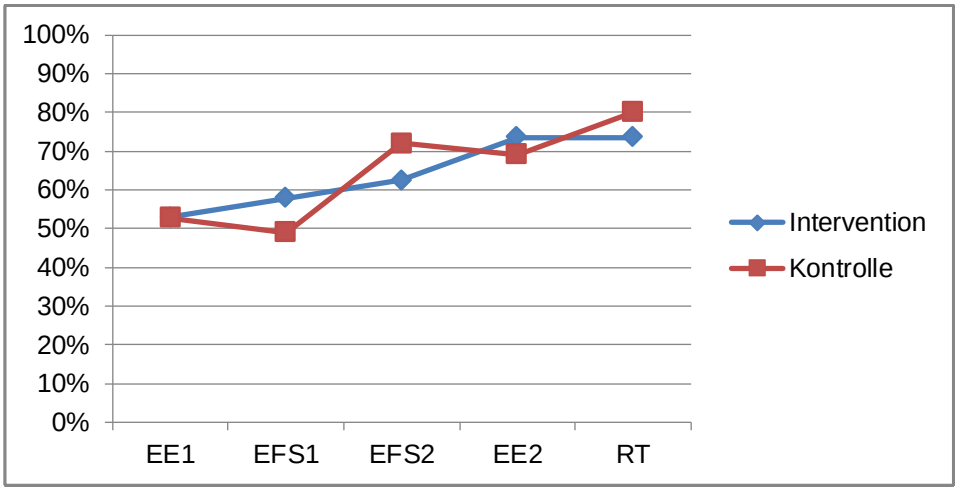
	Kipptisch-Untersuchung diagnostiziert werden kann.				
Erkennen einer Tachyarrhythmia absoluta im EKG	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Bitte erläutern Sie kurz, woran erkennbar ist, dass in diesem EKG am ehesten eine Tachyarrhythmia absoluta vorliegt.	40,52	29,04	0,8	0,76
Frequenzsenkung bei Tachyarrhythmia absoluta unbekannter Dauer	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Warum sollten Sie in dieser Situation keine Rhythmisierung (z.B. mittels elektrischer Kardioversion) anstreben?	39,28	44,09	0,68	0,67
laborchemische Diagnose einer manifesten Hyperthyreose nach Konstrastmittelgabe	Um Ihren Lernerfolg zu erhöhen, skizzieren Sie bitte kurz, warum hier eine Hyperthyreose und keine Hypothyreose vorliegt.	48,41	35,08	0,86	0,84
Absetzen von Amiodaron bei manifester Hyperthyreose	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Bitte erklären Sie in Stichpunkten, wieso in dieser Situation keine Thyreoidektomie empfohlen werden sollte.	54,52	55,14	0,56	0,59
klinischer Verdacht auf Lungenfibrose bei Dyspnoe und inspiratorischem Velcro-Knistern	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Um Ihren Lernerfolg zu erhöhen, skizzieren Sie bitte kurz, welche Indizien in diesem Fall für das Vorliegen einer Lungenfibrose und gegen das Vorliegen	37,13	40,77	0,53	0,39

	eines Lungenödems oder einer Pneumonie sprechen.				
Erkennen einer Restriktion in der Lungenfunktionsdiagnostik	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Um Ihren Lernerfolg zu erhöhen, begründen Sie bitte kurz, warum bei dem Verdacht auf eine restriktive Lungenerkrankung die Bestimmung der Vitalkapazität zur Diagnosestellung allein nicht ausreicht.	26,68	42,83	0,35	0,46
Indikationsstellung zur LTOT aufgrund eines pO2 <55 mmHg in der arteriellen BGA	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Um Ihren Lernerfolg zu erhöhen, erklären Sie bitte kurz den Unterschied zwischen einer Sauerstoff-Langzeit-Therapie und einer nicht-invasiven Beatmung.	38,28	44,13	0,34	0,32
Verdacht auf supraventrikuläre Genese einer Tachycardie anhand eines beschriebenen EKGs	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Skizzieren Sie kurz, was bei dem beschriebenen EKG gegen das Vorliegen einer Tachyarrhythmia absoluta spricht.	25,06	26,17	0,8	0,73
Adenosin zur Unterbrechung einer SVT	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Aus welchem Grund ist die Gabe von Adenosin in dieser Situation sinnvoller als die Gabe eines Betablockers oder eine Kardioversion?	52,47	53,32	0,59	0,44

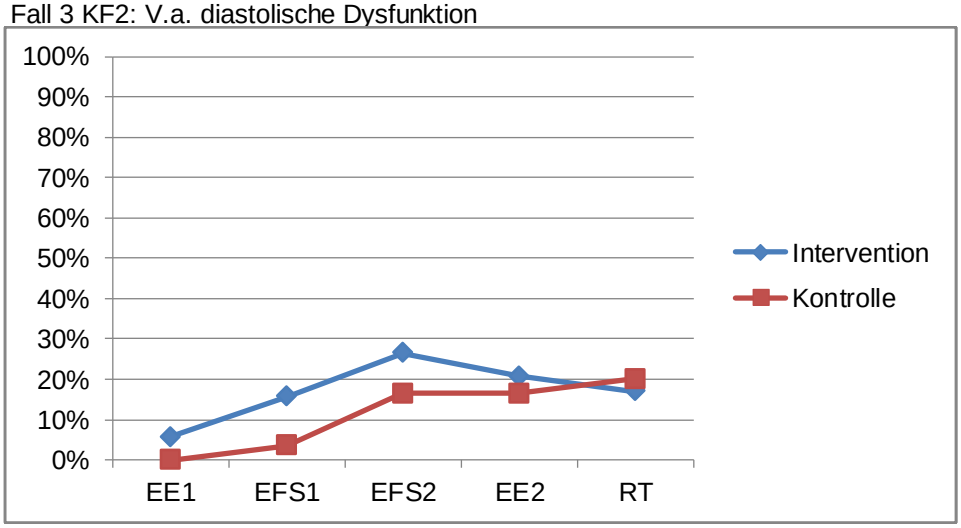
Elektrophysiologische Untersuchung zur Diagnosesicherung und Therapie	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Beschreiben Sie kurz, warum im vorliegenden Fall eher eine elektrophysiologische Untersuchung durchgeführt werden sollte, ohne vorher noch ein Langzeit-EKG anzufertigen.	55,12	53,25	0,52	0,46
Schrittmacher-Implantation bei kompletter AV-Blockierung	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Bitte erläutern Sie kurz, was ein 2-Kammer-Schrittmacher ist und wodurch er sich von einem biventrikulären Schrittmacher unterscheidet.	87,88	82,37	0,6	0,64
Verdacht auf Pneumothorax bei entsprechender Klinik nach Punktion der V. subclavia	Diese Frage wurde von Ihren Kommilitonen in der Vergangenheit häufig falsch beantwortet. Bitte erläutern Sie in Stichworten, was in diesem konkreten Fall für das Vorliegen eines Pneumothorax und gegen das Vorliegen eines Pleuraergusses spricht.	35,44	42,20	0,82	0,65

Anhang 6: Antworterfolg im Verlauf des Studienzeitraums je nach Item. X-Achse: Messzeitpunkte (EE1 = *Entry Exam*, EFS1 = E-Fallseminar 1, EFS2 = E-Fallseminar 2, EE2 = *Exit Exam*, RT = *Retention Test*). Y-Achse: Anteil richtiger Antworten. Diese Angaben gelten für alle der folgenden 30 Diagramme.

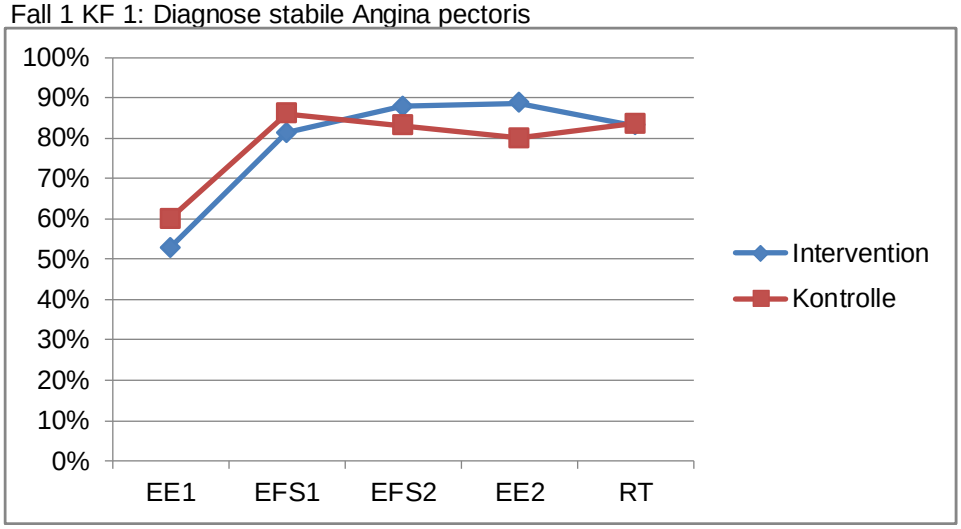
Fall 1 KF2: Ergometrie zum Ischämienachweis



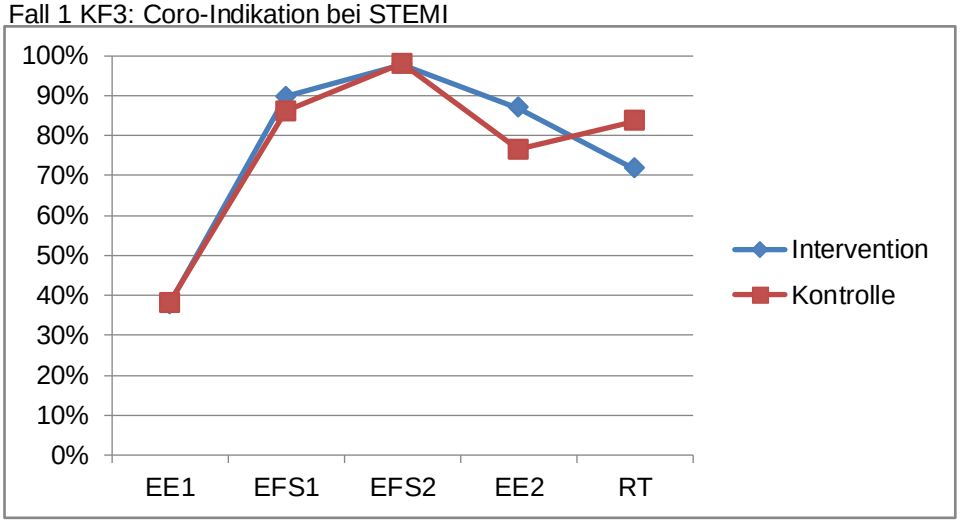
Fall 3 KF2: V.a. diastolische Dysfunktion



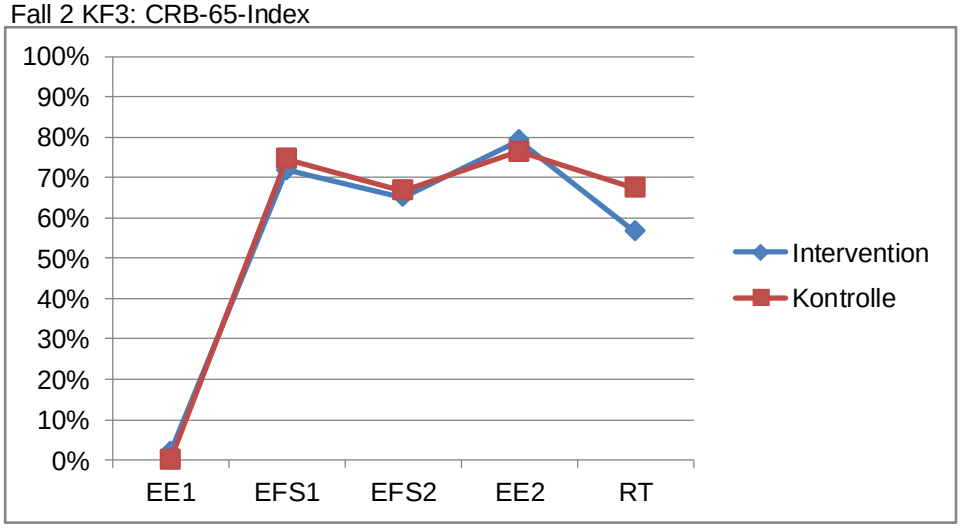
Fall 1 KF 1: Diagnose stabile Angina pectoris



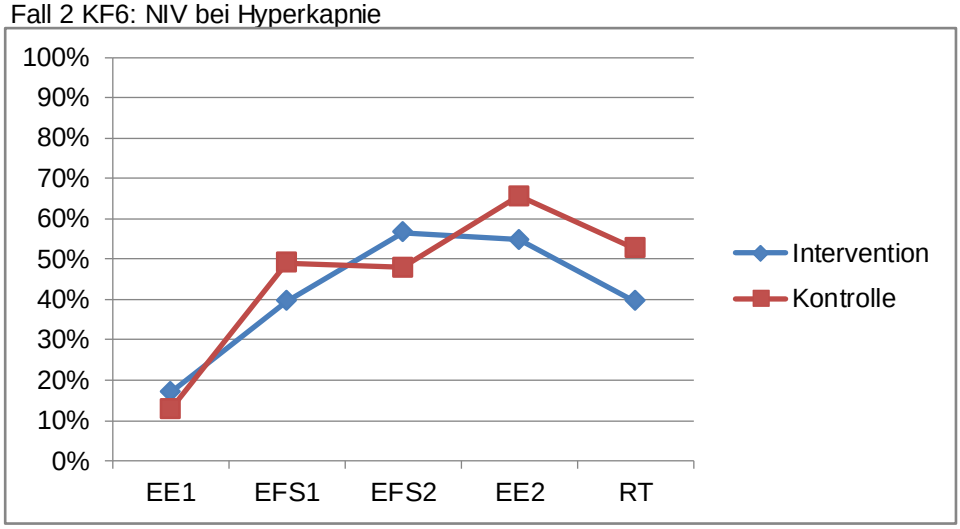
Fall 1 KF3: Coro-Indikation bei STEMI



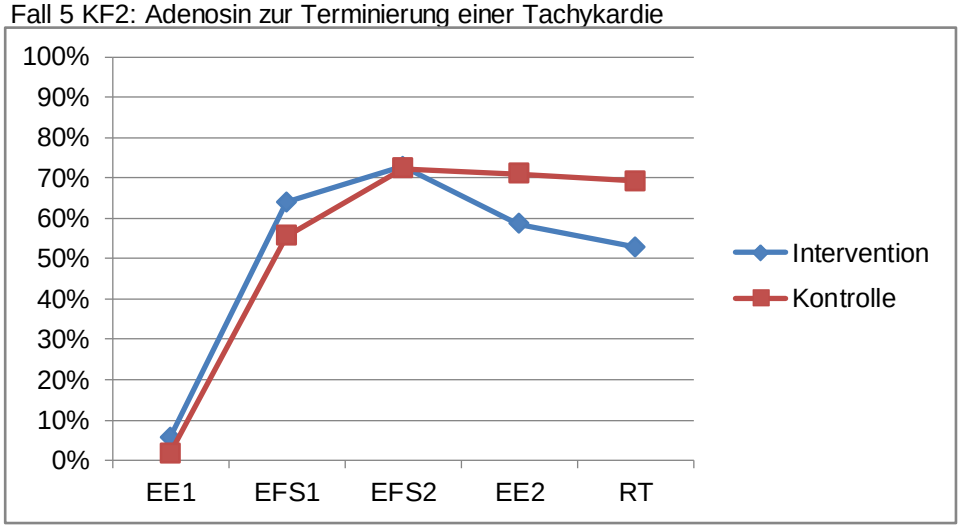
Fall 2 KF3: CRB-65-Index



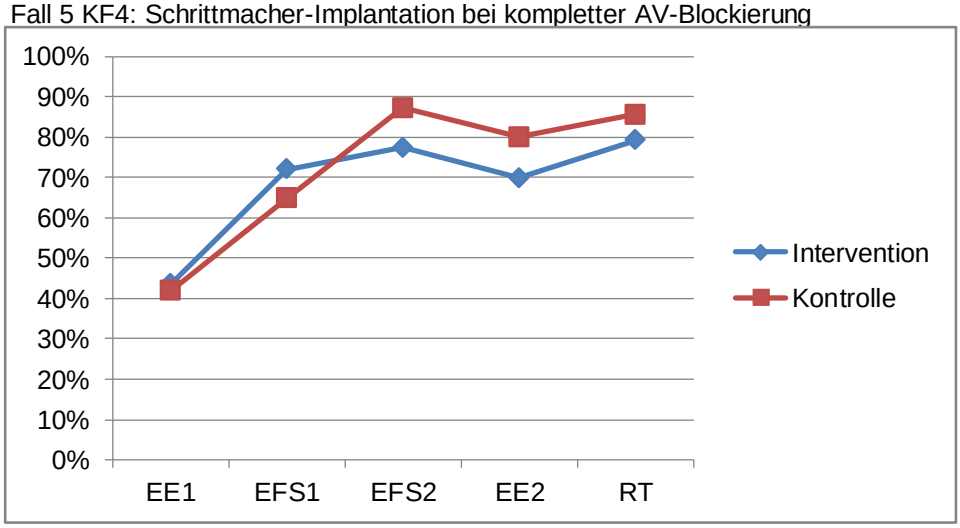
Fall 2 KF6: NIV bei Hyperkapnie



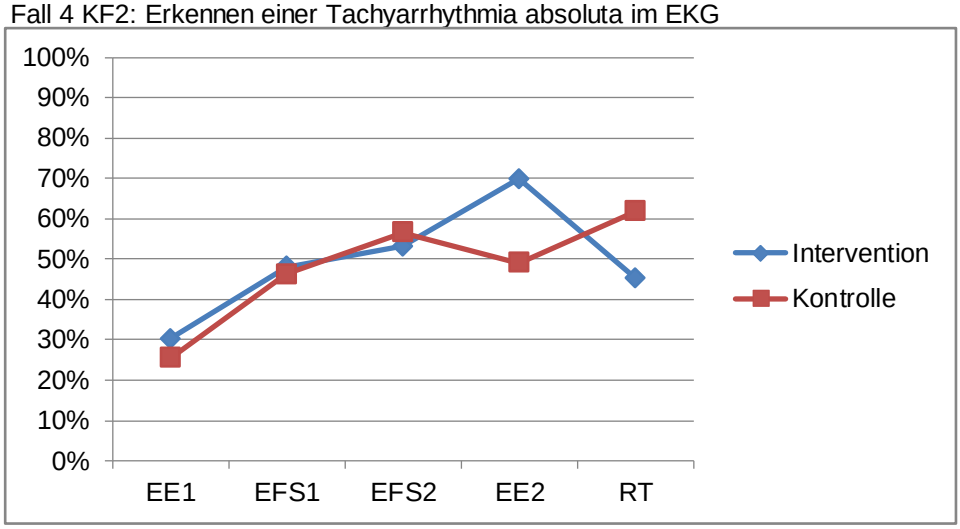
Fall 5 KF2: Adenosin zur Terminierung einer Tachykardie



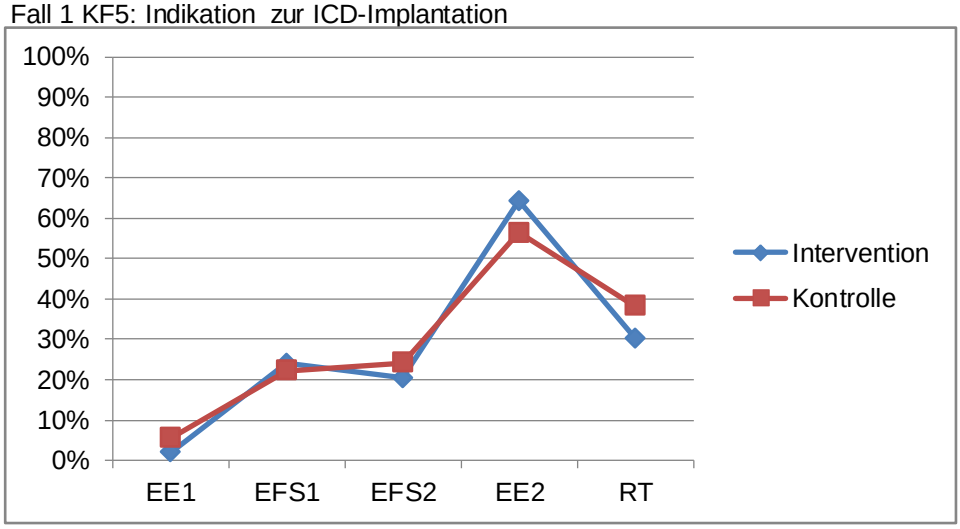
Fall 5 KF4: Schrittmacher-Implantation bei kompletter AV-Blockierung



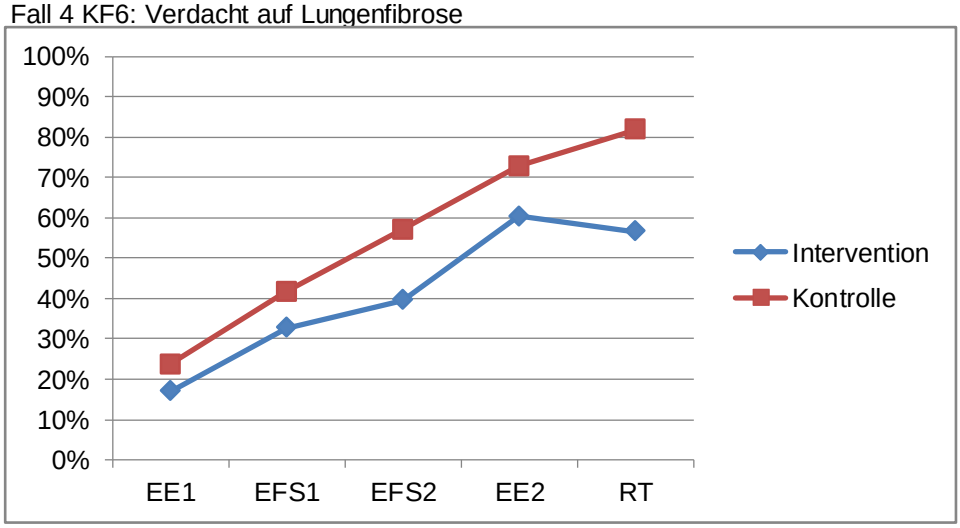
Fall 4 KF2: Erkennen einer Tachyarrhythmia absoluta im EKG



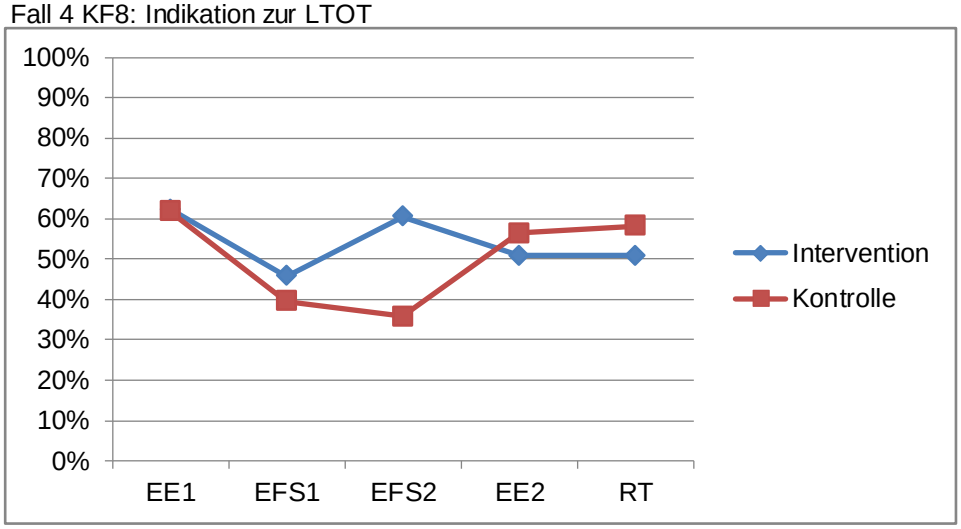
Fall 1 KF5: Indikation zur ICD-Implantation



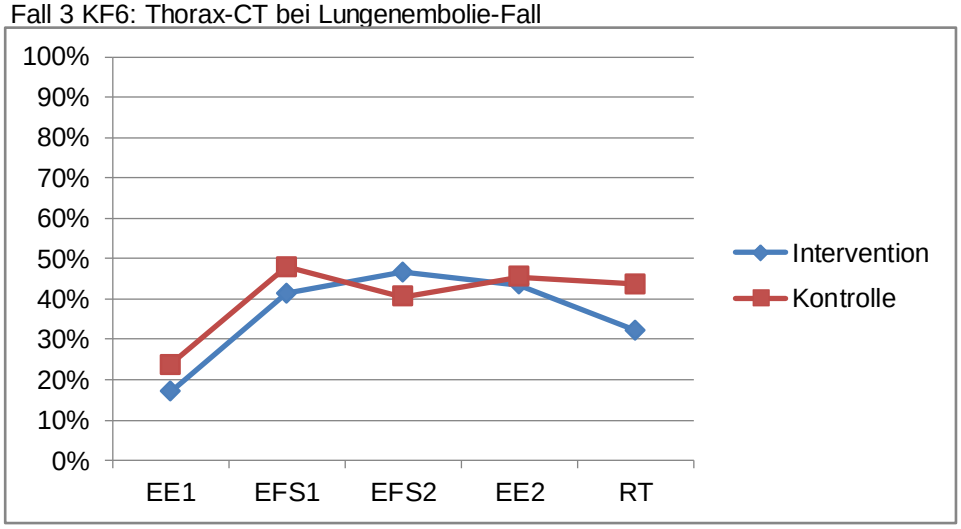
Fall 4 KF6: Verdacht auf Lungenfibrose



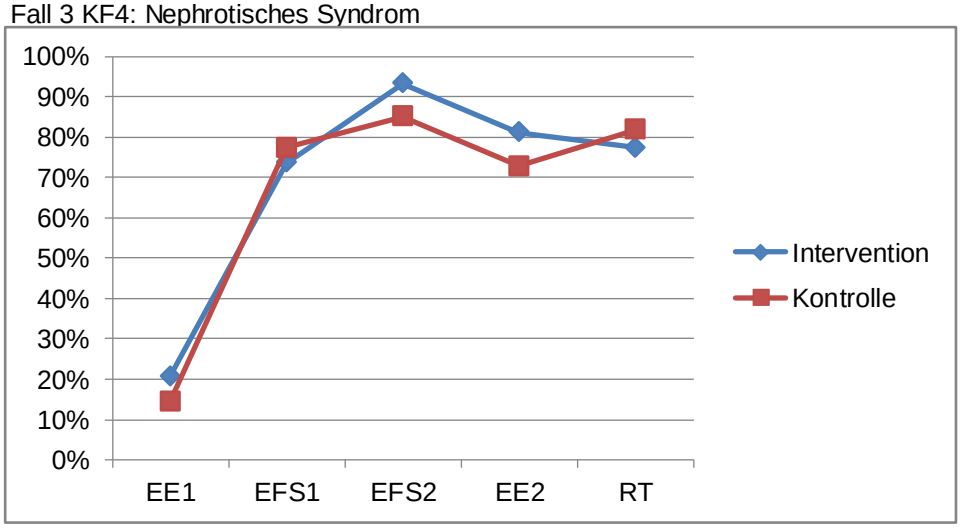
Fall 4 KF8: Indikation zur LTOT



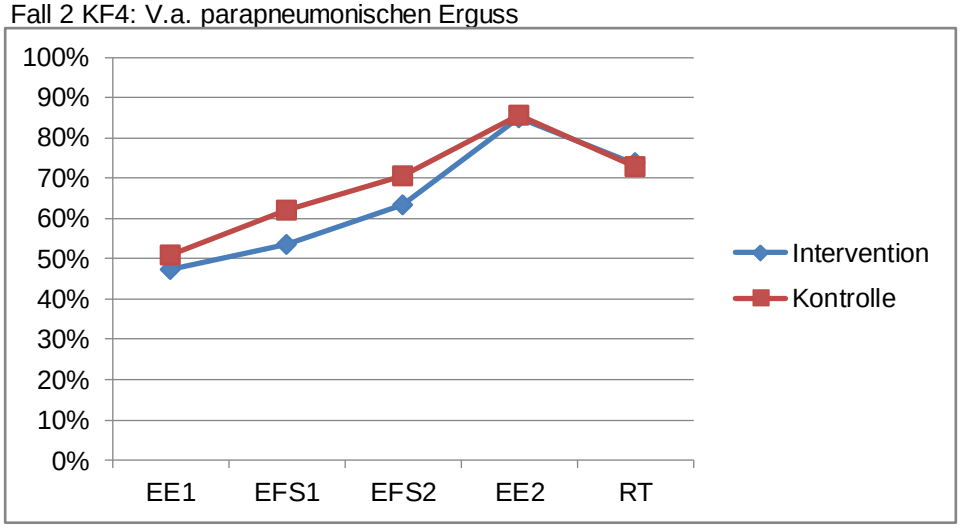
Fall 3 KF6: Thorax-CT bei Lungenembolie-Fall



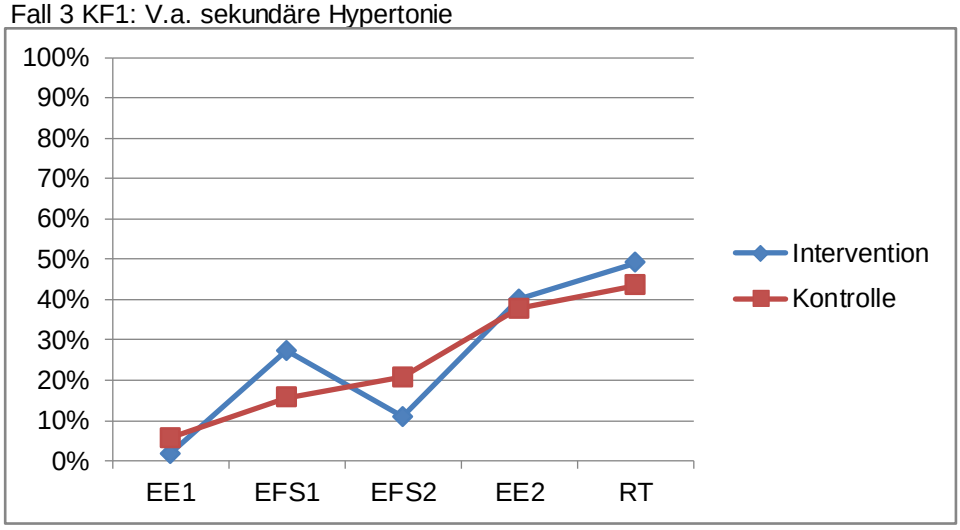
Fall 3 KF4: Nephrotisches Syndrom



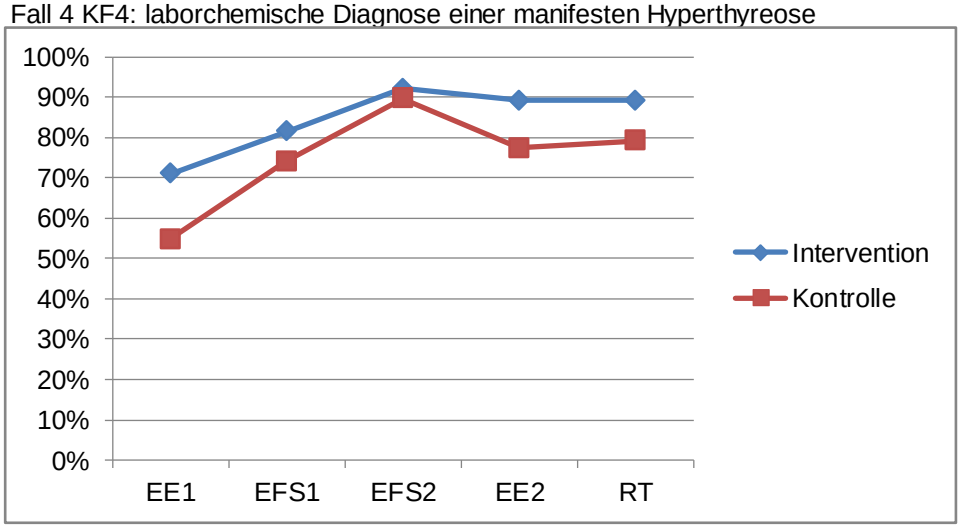
Fall 2 KF4: V.a. parapneumonischen Erguss



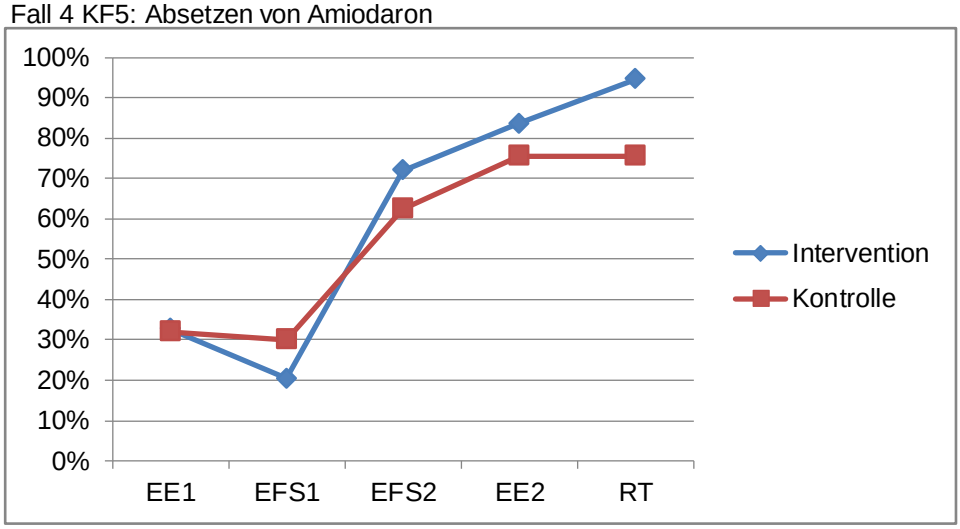
Fall 3 KF1: V.a. sekundäre Hypertonie



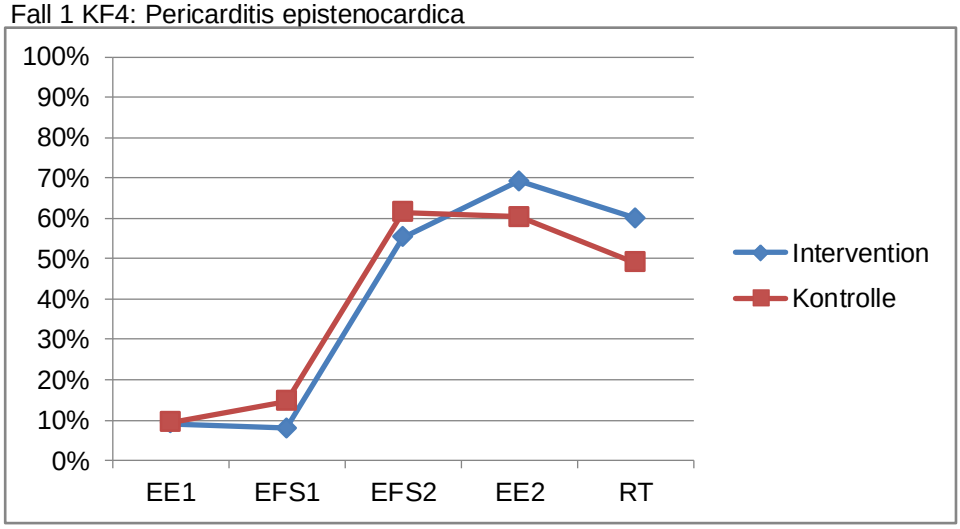
Fall 4 KF4: laborchemische Diagnose einer manifesten Hyperthyreose



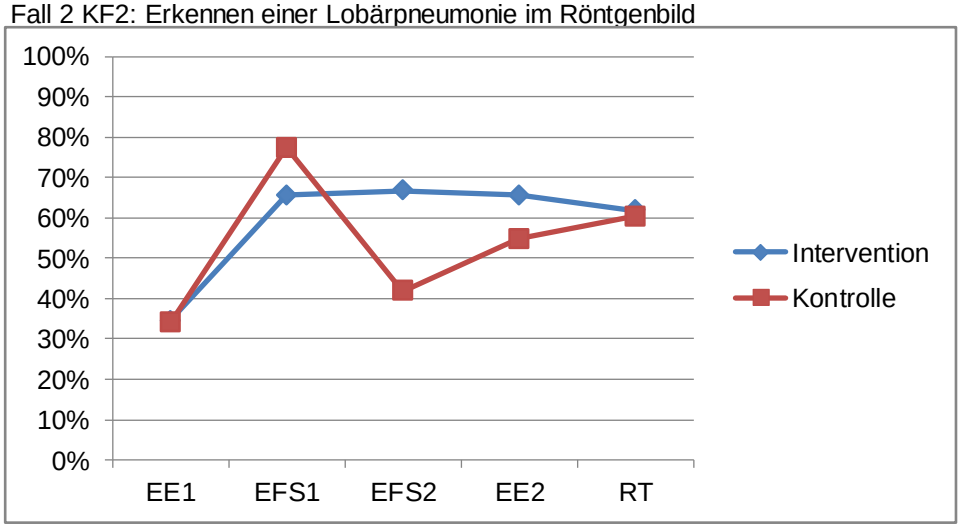
Fall 4 KF5: Absetzen von Amiodaron



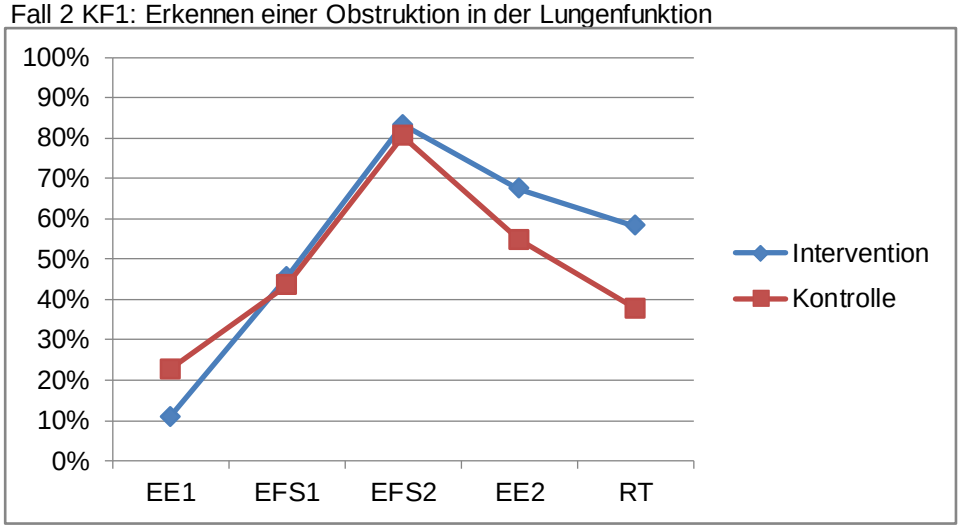
Fall 1 KF4: Pericarditis epistenocardica



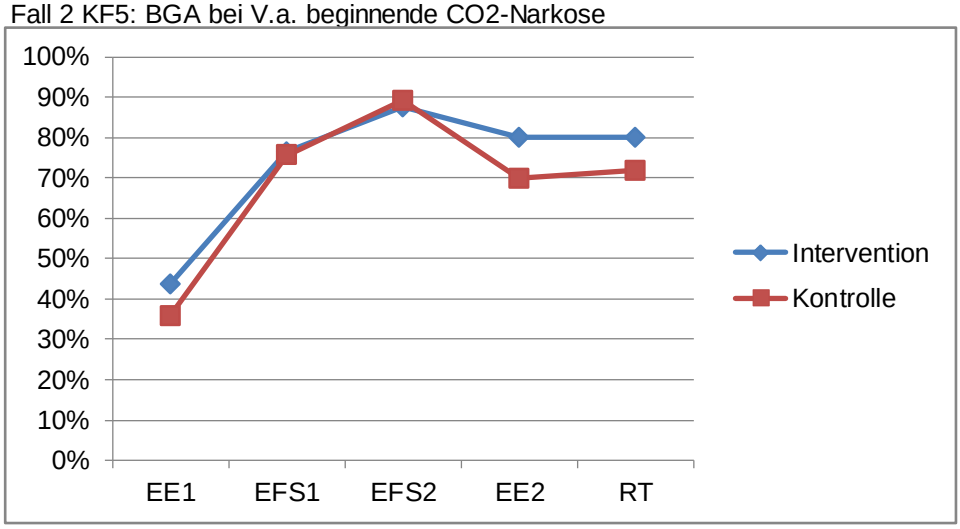
Fall 2 KF2: Erkennen einer Lobärpneumonie im Röntgenbild



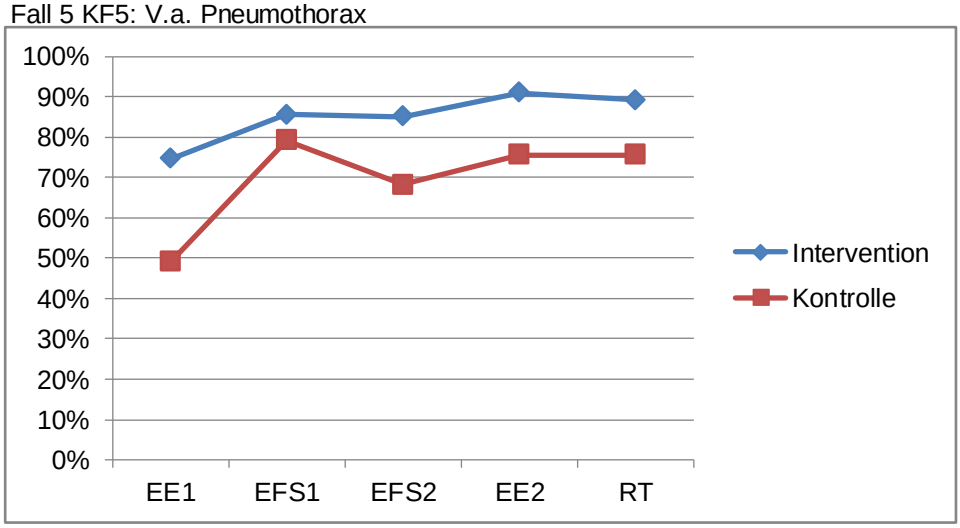
Fall 2 KF1: Erkennen einer Obstruktion in der Lungenfunktion



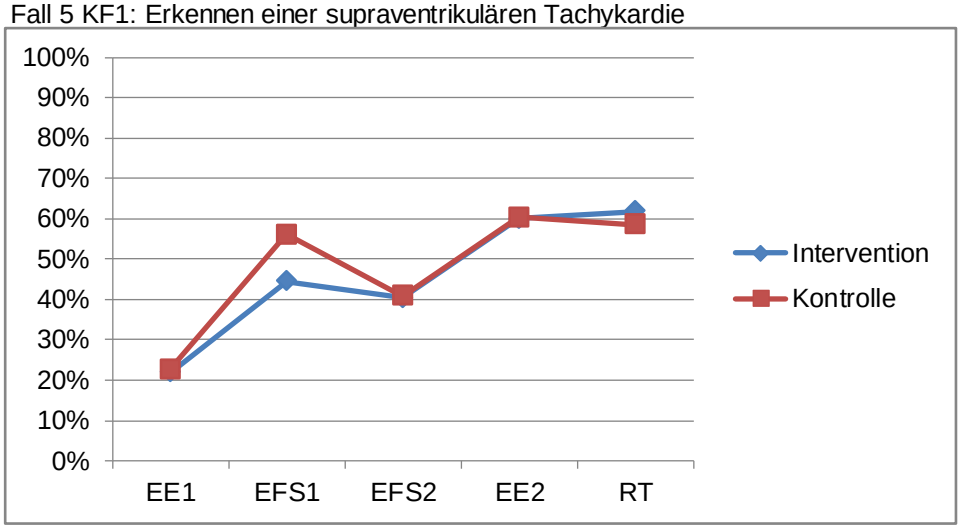
Fall 2 KF5: BGA bei V.a. beginnende CO2-Narkose



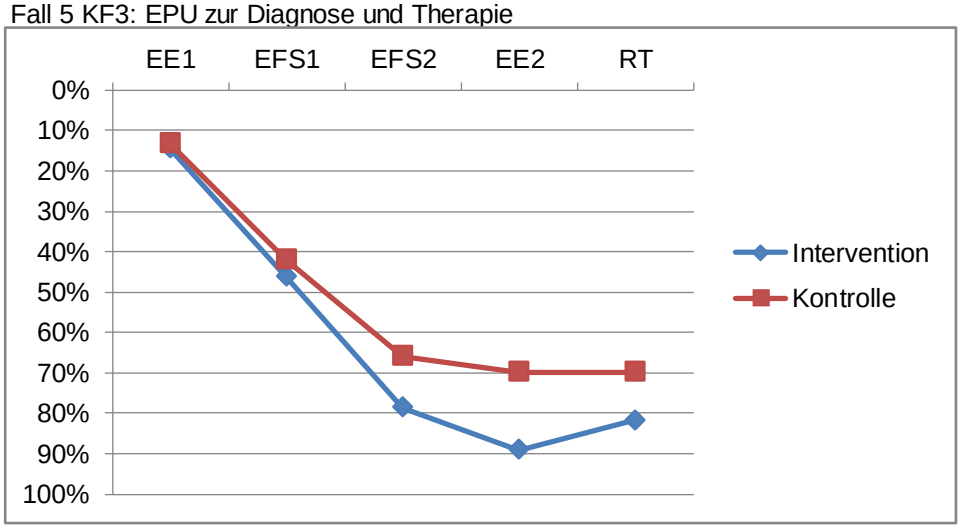
Fall 5 KF5: V.a. Pneumothorax



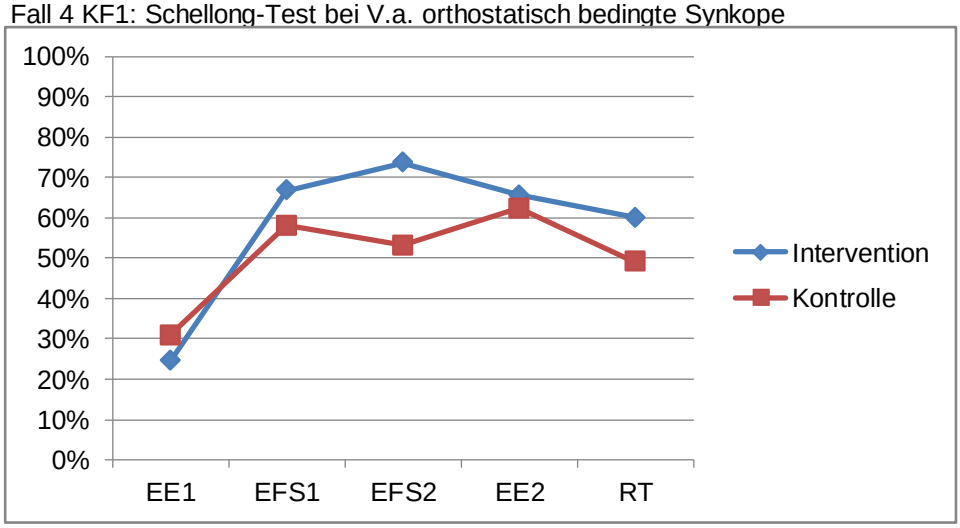
Fall 5 KF1: Erkennen einer supraventrikulären Tachykardie



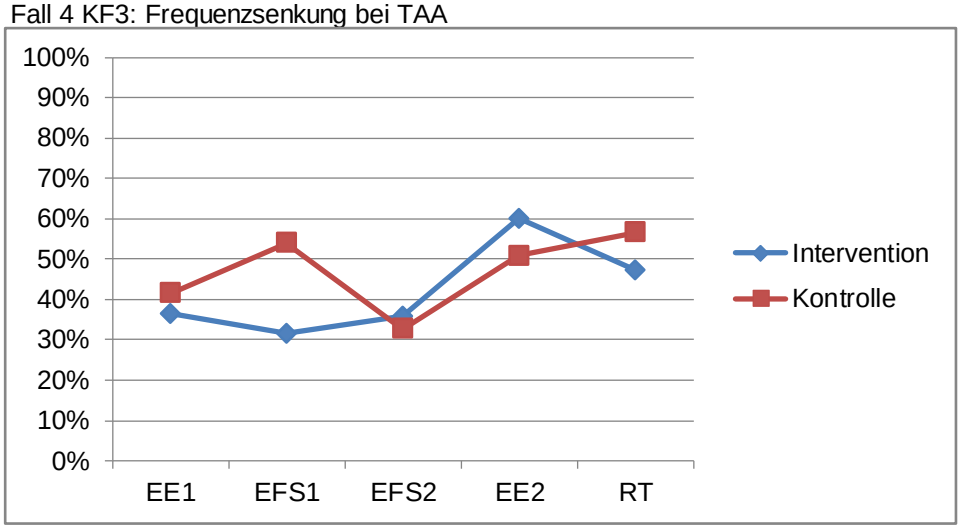
Fall 5 KF3: EPU zur Diagnose und Therapie



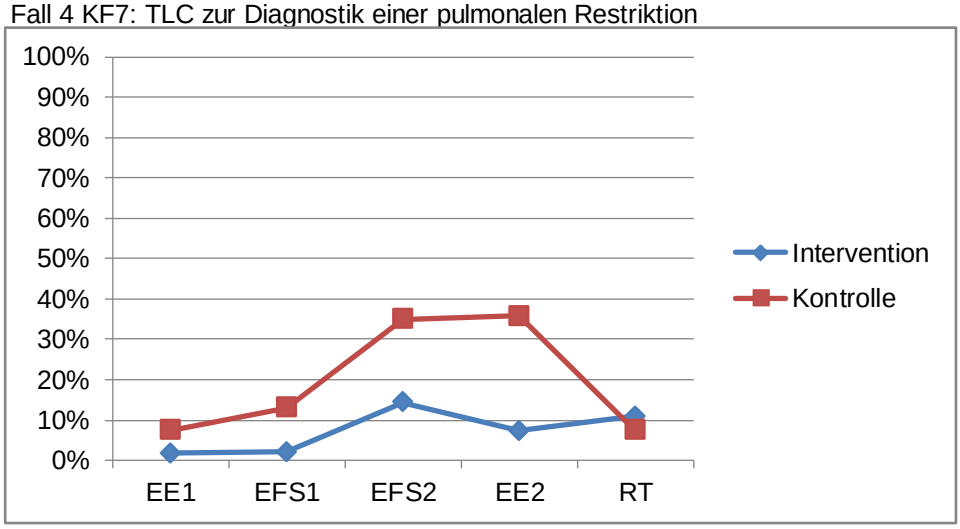
Fall 4 KF1: Schellong-Test bei V.a. orthostatisch bedingte Synkope



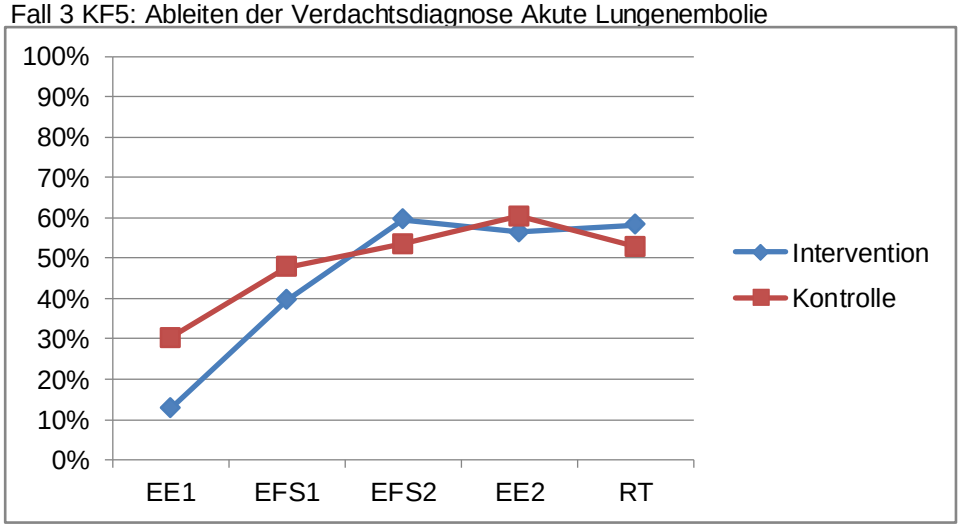
Fall 4 KF3: Frequenzsenkung bei TAA



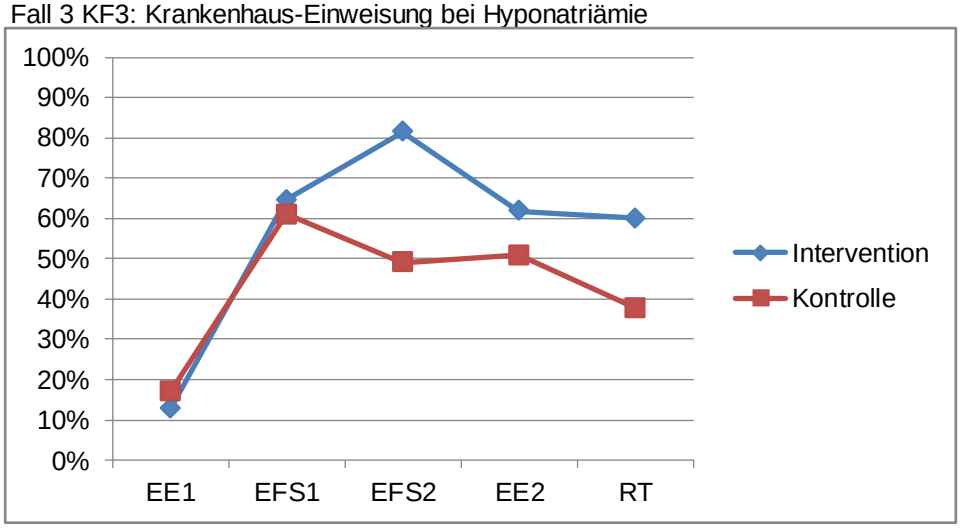
Fall 4 KF7: TLC zur Diagnostik einer pulmonalen Restriktion



Fall 3 KF5: Ableiten der Verdachtsdiagnose Akute Lungenembolie



Fall 3 KF3: Krankenhaus-Einweisung bei Hyponatriämie



7 Literaturverzeichnis

ÄApprO 2002: Approbationsordnung für Ärzte vom 27. Juni 2002 (BGBl. I S. 2405), die zuletzt durch Artikel 2 der Verordnung vom 22. September 2021 (BGBl. I S. 4335) geändert worden ist

Abatzis VT, Littlewood KE (2015): Debriefing in Simulation and Beyond. Int Anesthesiol Clin 53, 151–162

Bacon F: The new organon. Edited by Lisa Jardine and Michael Silverthorne (Cambridge texts in the history of philosophy). Cambridge University Press, Cambridge [U.K.] 2000

Beach MC, Inui T, The Relationship-Centered Care Research Network (2006): Relationship-centered Care: A Constructive Reframing. J Gen Intern Med 21, 3–8

Berufsverband deutscher Internisten e.V.: Innere Medizin – Der Internist. URL: https://www.internisten-im-netz.de/fachgebiete/innere-medizinder-internist.html; Zugriff am 10.04.2020

Biggs J (1996): Enhancing teaching through Constructive Alignment. High Educ 32, 347–364

Bordage G, Page G: An alternative to PMPs: The "Key Feature concept". In: Hart IR, Harden R (Hrsg.): Further developments in assessing clinical competence. Can-Heal Publications, Ottawa 1987, 59–75

Bordage G, Lemieux M (1991): Semantic structures and diagnostic thinking of experts and novices. Acad Med 66, 70–2

Bordage G, Page G (2018): The Key Features approach to assess clinical decisions: validity evidence to date. Adv Health Sci Educ 23, 1005–1036

Bowen JL (2006): Educational Strategies to Promote Clinical Diagnostic Reasoning. N Engl J Med 355, 2217–2225

Bruin ABH de, Schmidt HG, Rikers RMJP (2005): The Role of Basic Science Knowledge and Clinical Knowledge in Diagnostic Reasoning: A Structural Equation Modeling Approach. Acad Med 80, 765–773

Bundesärztekammer 2019: (Muster-)Berufsordnung für die in Deutschland tätigen Ärztinnen und Ärzte in der Fassung der Beschlüsse des 121. Deutschen Ärztetages 2018 in Erfurt, geändert durch Beschluss des Vorstandes der Bundesärztekammer am 14.12.2018.

Butler AC, Roediger HL (2007): Testing improves long-term retention in a simulated classroom setting. Eur J Cogn Psychol 19, 514–527

Cerutti B, Blondon K, Galetto A (2016): Long Menu questions in computer-based assessments: a retrospective observational study. BMC Med Educ 16, 55

Chamberland M, Mamede S, Bergeron L, Varpio L (2020): A layered analysis of Self-explanation and structured reflection to support Clinical Reasoning in medical students. Perspect Med Educ 10, 171–179

Chang RW, Bordage G, Connell KJ (1998): The importance of early problem representation during case presentations. Acad Med 73, 109-111

Charlin B, Lubarsky S, Millette B, Crevier F, Audétat M-C, Charbonneau A, Caire Fon N, Hoff L, Bourdy C (2012): Clinical Reasoning processes: unravelling complexity through graphical representation. Med Educ 46, 454–463

Cortina JM (1993): What is coefficient alpha? An examination of theory and applications. J Appl Psychol 78, 98–104

Croskerry P (2000): The Cognitive Imperative: Thinking about How We Think. Acad Emerg Med 7, 1223–1231

Croskerry P (2003): Cognitive forcing strategies in clinical decisionmaking. Ann Emerg Med 41, 110–120

Croskerry P (2009): A Universal Model of Diagnostic Reasoning. Acad Med 84, 1022–1028

Cutrer WB, Sullivan WM, Fleming AE (2013): Educational Strategies for Improving Clinical Reasoning. Curr Probl Pediatr Adolesc Health Care 43, 248–257

Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, Ratcliffe T, Gordon D, Heist B, Lubarsky S, et al. (2019): Clinical Reasoning Assessment Methods: A Scoping Review and Practical Guidance. Acad Med 94, 902–912

Dempsey JV: Interactive Instruction and Feedback. Educational Technology Publications, Englewood Cliffs 1993

Dobson J, Linderholm T, Perez J (2018): Retrieval practice enhances the ability to evaluate complex physiology information. Med Educ 52, 513–525

Elstein AS (2009): Thinking about diagnostic thinking: a 30-year perspective. Adv Health Sci Educ 14, 7–18

Elstein AS, Shulman, LS, Sprafka SA: Medical Problem Solving: an Analysis of Clinical Reasoning. Harv Univ Press, Cambridge [Mass.] 1978

Eva KW (2005): What every teacher needs to know about Clinical Reasoning. Med Educ 39, 98–106

Eva KW, Wood TJ (2003): Can the Strength of Candidates Be Discriminated Based on Ability to Circumvent the Biasing Effect of Prose? Implications for Evaluation and Education. Acad Med 78, 78-81

Eva KW, Wood TJ, Riddle J, Touchie C, Bordage G (2010): How clinical features are presented matters to weaker diagnosticians. Med Educ 44, 775–785

Faller H, Lang H: Medizinische Psychologie und Soziologie. Springer-Verlag, Berlin 2011

Farmer EA, Page G (2005): A practical guide to assessing clinical decision-making skills using the Key Features approach. Med Educ 39, 1188–1194

Fischer MR, Kopp V, Holzer M, Ruderich F, Jünger J (2005): A modified electronic Key Feature examination for undergraduate medical students: validation threats and opportunities. Med Teach 27, 450–455

Fitch ML, Drucker AJ, Norton JAJr (1951): Frequent testing as a motivating factor in large lecture classes. J Educ Psychol 42, 1–20

Fölsch UR, Hallek M, Raupach T, Hasenfuß G (2017): Resonanz und Weiterentwicklung der Initiative Klug entscheiden. Internist 58, 527–531

Fournier C von (2000): Zeitmanagement: Fest umrissene Ziele und klare Prioritäten. Dtsch Arztebl Int 97, A-2374-2377

Gartmeier M, Bauer J, Gruber H, Heid H (2008): Negative Knowledge: Understanding Professional Learning and Expertise. Vocat Learn 1, 87–103

Goldmann M, Hasenfuß G, Dehl T, Raupach T (2016): Klug entscheiden: . . . auch in der Lehre! Dtsch Arztebl Int 113, A-2149-2154

Goldmann M, Middeke A-C, Schuelper N, Dehl T, Raupach T (2017): Klug entscheiden in der Lehre. Z Evid Fortbild Qual Gesundhwes 129, 22–26

Graber ML, Franklin N, Gordon R (2005): Diagnostic Error in Internal Medicine. Arch Intern Med 165, 1493

Grant J, Marsden P (1987): The structure of memorized knowledge in students and clinicians: an explanation for diagnostic expertise. Med Educ 21, 92–98

Grant J, Marsden P (1988): Primary knowledge, medical education and consultant expertise. Med Educ 22, 173–179

Green ML, Moeller JJ, Spak JM (2018): Test-Enhanced Learning in health professions education: A systematic review: BEME Guide No. 48. Med Teach 40, 337–350

Groves M (2011): Fostering Clinical Reasoning in medical students. Med Educ 45, 518–519

Groves M (2012): Understanding Clinical Reasoning: the next step in working out how it really works. Med Educ 46, 444–446

Gruppen LD (2017): Clinical Reasoning: Defining It, Teaching It, Assessing It, Studying It. West J Emerg Med 18, 4

Hall KH (2002): Reviewing intuitive decision-making and uncertainty: the implications for medical education. Med Educ 36, 216–224

Hasselhorn M, Gold A: Pädagogische Psychologie: erfolgreiches Lernen und Lehren. W. Kohlhammer Verlag, Stuttgart 2009

Hatala R, Norman GR (2002): Adapting the Key Features Examination for a clinical clerkship. Med Educ 36, 160–165

Heitzmann N, Fischer F, Kühne-Eversmann L, Fischer MR (2015): Enhancing diagnostic competence with Self-explanation prompts and adaptable feedback. Med Educ 49, 993–1003

Hrynchak P, Glover Takahashi S, Nayer M (2014): Key Feature questions for assessment of Clinical Reasoning: a literature review. Med Educ 48, 870–883

Huwendiek S, Reichert F, Duncker C, de Leng BA, van der Vleuten CPM, Muijtjens AMM, Bosse H-M, Haag M, Hoffmann GF, Tönshoff B, Dolmans D (2017): Electronic assessment of Clinical Reasoning in clerkships: A mixed-methods comparison of Long Menu Key Feature problems with context-rich single best answer questions. Med Teach 39, 476–485

James W: The principles of psychology. Holt, New York 1890

Kahneman D: Thinking, fast and slow. 1. Auflage; Farrar, Straus and Giroux, New York 2011

Kassirer JP (2010): Teaching Clinical Reasoning: Case-Based and Coached. Acad Med 85, 1118–1124

Kromann CB, Jensen ML, Ringsted C (2009): The effect of testing on skills learning. Med Educ 43, 21–27

Larsen DP, Butler AC, Roediger HL (2008): Test-Enhanced Learning in medical education. Med Educ 42, 959–966

Larsen DP, Butler AC, Roediger HL (2009): Repeated testing improves long-term retention relative to repeated study: a randomised controlled trial. Med Educ 43, 1174–1181

Larsen DP, Butler AC, Roediger HL (2013): Comparative effects of Test-Enhanced Learning and Self-explanation on long-term retention. Med Educ 47, 674–682

Levin JR (1988): Elaboration-based learning strategies: Powerful theory = powerful application. Contemp Educ Psychol 13, 191–205

Levine D, Bleakley A (2013): Rethinking Clinical Reasoning. Med Educ 47, 745–746

Ludwig S, Schuelper N, Brown J, Anders S, Raupach T (2018): How can we teach medical students to choose wisely? A randomised controlled cross-over study of video- versus text-based case scenarios. BMC Med 16, 107

Mann K, Gordon J, MacLeod A (2009): Reflection and reflective practice in health professions education: a systematic review. Adv Health Sci Educ 14, 595–621

McDaniel MA, Roediger HL, McDermott KB (2007): Generalizing Test-Enhanced Learning from the laboratory to the classroom. Psychon Bull Rev 14, 200–206

MFT Medizinischer Fakultätentag der Bundesrepublik Deutschland e. V. (2015): Nationaler Kompetenzbasierter Lernzielkatalog Medizin (NKLM). http://www.nklm.de/kataloge/nklm/lernziel/uebersicht; abgerufen am 22.02.2020

Modi JN, Anshu, Gupta P, Singh T (2015): Teaching and assessing Clinical Reasoning skills. Indian Pediatr 52, 787–794

Muzzin L, Norman G, Feightner J, Tugwell P: Expertise in recall of clinical protocols in two specialty areas. In: Proceedings of the 22nd Conference on Research in Medical Education. Band 22; Association of American Medical Colleges, Washington [D.C.] 1983, 122–127

Neufeld VR, Norman GR, Feightner JW, Barrows HS (1981): Clinical problem-solving by medical students: a cross-sectional and longitudinal analysis. Med Educ 15, 315–322

Newble DI, Jaeger K (1983): The effect of assessments and examinations on the learning of medical students. Med Educ 17, 165–171

Nikendei C, Mennin S, Weyrich P, Kraus B, Zipfel S, Schrauth M, Jünger J (2009): Effects of a supplementary final year curriculum on students' Clinical Reasoning skills as assessed by Key Feature examination. Med Teach 31, e438–e442

Norman G (2005): Research in Clinical Reasoning: past history and current trends. Med Educ 39, 418–427

Norman G (2012): Medical education: past, present and future. Perspect Med Educ 1, 6–14

Norman GR, Eva KW (2010): Diagnostic error and Clinical Reasoning. Med Educ 44, 94–100

Norman G, Bordage G, Curry L, Dauphinee D, Jolly B, Newble D, et al.: Review of recent innovations in assessment. In: Wakeford R (Hrsg.): Directions in Clinical Assessment: Report of the Cambridge Conference on the Assessment of Clinical Competence. Cambridge University School of Clinical Medicine, Cambridge [U.K.] 1984, 9–27

Norman GR, Tugwell P, Feightner JW, Muzzin LJ, Jacoby LL (1985): Knowledge and clinical problem-solving. Med Educ 19, 344–356

Norman G, Bordage G, Page G, Keane D (2006): How specific is case specificity? Med Educ 40, 618–623

Nungester RJ, Duchastel PC (1982): Testing versus review: Effects on retention. J Educ Psychol 74, 18–22

Page G, Bordage G (1995): The Medical Council of Canada's Key Features project: a more valid written examination of clinical decision-making skills. Acad Med 70, 104-110

Page G, Bordage G, Allen T (1995): Developing Key Feature problems and examinations to assess clinical decision-making skills. Acad Med 70, 194–201

Papa FJ, Shores JH, Meyer S (1990): Effects of pattern matching, pattern discrimination, and experience in the development of diagnostic expertise. Acad Med 65, S21-S22

Patel VL, Groen GJ, Frederiksen CH (1986): Differences between medical students and doctors in memory for clinical cases. Med Educ 20, 3–9

Patel VL, Kaufman DR, Arocha JF (2002): Emerging paradigms of cognition in medical decision-making. J Biomed Inform 35, 52–75

Pinnock R, Welch P (2014): Learning Clinical Reasoning. J Paediatr Child Health 50, 253–257

Raupach T, Schuelper N (2018): Reconsidering the role of assessments in undergraduate medical education. Med Educ 52, 464–466

Raupach T, Andresen JC, Meyer K, Strobel L, Koziolek M, Jung W, Brown J, Anders S (2016): Test-Enhanced Learning of Clinical Reasoning: a crossover randomised trial. Med Educ 50, 711–720

Redelmeier DA (2005): The Cognitive Psychology of Missed Diagnoses. Ann Intern Med 142, 115

Rikers RMJP, Schmidt HG, Boshuizen HPA (2000): Knowledge Encapsulation and the Intermediate Effect. Contemp Educ Psychol 25, 150–166

Roediger HL, Karpicke JD (2006a): Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. Psychol Sci 17, 249–255

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

Rotthoff T, Baehring T, Dicken H-D, Fahron U, Richter B, Fischer MR, Scherbaum WA (2006): Comparison between Long Menu and Open-Ended Questions in computerized medical assessments: A randomized controlled trial. BMC Med Educ 6, 50

Schmidmaier R, Ebersbach R, Schiller M, Hege I, Holzer M, Fischer MR (2011): Using electronic flashcards to promote learning in medical students: retesting versus restudying. Med Educ 45, 1101–1110

Schmidt HG, Boshuizen HPA (1993): On the origin of intermediate effects in clinical case recall. Mem Cognit 21, 338–351

Schmidt HG, Mamede S (2015): How to improve the teaching of Clinical Reasoning: a narrative review and a proposal. Med Educ 49, 961–973

Schmidt HG, Norman GR, Boshuizen HP (1990): A cognitive perspective on medical expertise: theory and implication [published erratum appears in Acad Med 1992 Apr;67(4):287]. Acad Med 65, 611–621

Schnabel KP, Boldt PD, Breuer G, Fichtner A, Karsten G, Kujumdshiev S, Schmidts M, Stosch C (2011): Konsensusstatement „Praktische Fertigkeiten im Medizinstudium" – ein Positionspapier des GMA-Ausschusses für praktische Fertigkeiten. GMS Z Med Ausbild 28, Doc58

Schuwirth LWT: An approach to the assessment of medical problem solving: Computerised case-based testing. Doc. Phil. Diss. Maastricht 1998

Schuwirth LW, van der Vleuten CP, Stoffers HE, Peperkamp AG (1996): Computerized Long Menu questions as an alternative to open-ended questions in computerized assessment. Med Educ 30, 50–55

Slavin RE (1996): Research on Cooperative Learning and Achievement: What We Know, What We Need to Know. Contemp Educ Psychol 21, 43–69

Stark R, Kopp V, Fischer MR (2011): Case-based learning with worked examples in complex domains: Two experimental studies in undergraduate medical education. Learn Instr 21, 22–33

Sullivan ME, Park YS, Liscum K, Sachdeva AKM, Gesbeck M, Bordage G (2020): The American College of Surgeons Entering Resident Readiness Assessment Program: Development and National Pilot Testing Results. Ann Surg 272, 194-198

VanLehn K (1999): Rule-Learning Events in the Acquisition of a Complex Skill: An Evaluation of Cascade. J Learn Sci 8, 71–125

Weingardt M (2014): Wer aufhört Fehler zu machen, lernt nicht mehr dazu. Lernen und Lernstörungen 3, 23–38

Weinstein CE, Jung J, Acee TW: Learning Strategies. In: Aukrust VG (Hrsg.): Learning and Cognition. Elsevier, Oxford 2011, 137–143

Wood T (2009): Assessment not only drives learning, it may also help learning. Med Educ 43, 5–6

Danksagung

Ich bedanke mich herzlich bei meinem Doktorvater Prof. Dr. med. Tobias Raupach für seine Unterstützung mit nie nachlassendem Engagement, für seinen Elan und seine Kreativität, für seinen Humor und seine Freundlichkeit, und nicht zuletzt für die langanhaltende Geduld.