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

Effektivität videogestützter *Key-Feature*-Prüfungen beim Erwerb differentialdiagnostischer und -therapeutischer Kompetenzen im Medizinstudium

INAUGURAL-DISSERTATION

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

vorgelegt von

Sascha Ludwig

aus

Bielefeld

Göttingen 2021

Dekan:	Prof. Dr. med. W. Brück
Referent:	Prof. Dr. med. T. Raupach, MME
Ko-Referentin:	PD Dr. med. S. Sennhenn-Kirchner, MME

Tag der mündlichen Prüfung: 17.05.2022

Hiermit erkläre ich, die Dissertation mit dem Titel „Effektivität videogestützter *Key-Feature*-Prüfungen beim Erwerb differentialdiagnostischer und -therapeutischer Kompetenzen im Medizinstudium“ eigenständig angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Göttingen, den 17.05.2021 _____

(Unterschrift)

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

Ludwig S, Meyer K, Anders S, Raupach T: Test-enhanced learning with key feature questions: Video-based exams are more effective than text-based exams (28098). *Short communication* (#9F4) im Rahmen der Association for Medical Education in Europe (AMEE) 2015, Glasgow 04.09.-09.09.2015

Ludwig S, Meyer K, Schuelper N, Anders S, Raupach T: Effektivität videobasierter *Key-Feature*-Prüfungen im Vergleich zu textbasierten Prüfungen in der Inneren Medizin (DocV445). In: Haak R und Fuchß A (Hrsg.): Gemeinsame Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA) und des Arbeitskreises zur Weiterentwicklung der Lehre in der Zahnmedizin (AKWLZ) 30.09.-03.10.2015 in Leipzig; German Medical Science GMS Publishing House, Düsseldorf 2015

Ludwig S, Meyer K, Anders S, Raupach T: Test-enhanced learning with key features: video- versus text-based questions (133527). *Short communication* (#7P2) im Rahmen der Association for Medical Education in Europe (AMEE) 2016, Barcelona 27.08.-31.08.2016

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

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

ÄApprO	Approbationsordnung für Ärzte
ACE	*Angiotensin Converting Enzyme*
AK	Antikörper
BGA	Blutgasanalyse
BNP	*Brain Natriuretic Peptide*
bvmd	Bundesvertretung der Medizinstudierenden in Deutschland e. V.
CBL	*Case-based learning*
COPD	*Chronic obstructive pulmonary disease*
CT	Computertomografie
DiPS	Digitales Prüfungs- und Schulungszentrum
EF	Ejektionsfraktion
EKG	Elektrokardiogramm
GMA	Gesellschaft für medizinische Ausbildung
IMS	*Item Management System*
KHK	Koronare Herzerkrankung
LM	*Long Menu*
LTOT	*Long term oxygen therapy*
MC	*Multiple Choice*
NIV	Nichtinvasive Ventilation
NKLM	Nationaler Kompetenzbasierter Lernzielkatalog Medizin
NSTEMI	*Non-ST-elevation myocardial infarction*
OSCE	*Objective structured clinical examination*
PBL	*Problem-based learning*
PMP	*Patient Management Problem*
POL	Problemorientiertes Lernen
POLEMA	Problemorientiertes Lernen, elektronische Medien assistiert
QTI-Datei	*Question & Test Interoperability*-Datei
SLE	Systemischer Lupus erythematodes
SPs	Standardisierte Patienten
STÄPS	Studentisches Trainingszentrum ärztlicher Praxis und Simulation
STEMI	*ST-elevation myocardial infarction*
UaK	Unterricht am Krankenbett
UMG	Universitätsmedizin Göttingen

1 Einleitung

Eine der Schlüsselkompetenzen von Ärzten[1] ist das sogenannte „klinische Denken“, die klinische Entscheidungsfindung bezüglich diagnostischer und therapeutischer Möglichkeiten in der Patientenversorgung. Neben der Vermittlung von faktischem und konzeptionellem Wissen sollte demnach auch die Vermittlung dieser prozeduralen Kompetenzen als Lernziel in der Lehre und Ausbildung zukünftiger Ärzte einen hohen Stellenwert einnehmen.

Allgemein lassen sich für Lernziele nach Blooms überarbeiteter Taxonomie folgende drei Domänen unterscheiden: die kognitive (Wissen), die affektive (Einstellungen) und die psychomotorische (praktische Fähigkeiten) Domäne. Die kognitive Domäne lässt sich weiter unterteilen in faktisches, konzeptionelles, prozedurales und metakognitives Wissen (Krathwohl 2002). In der vorliegenden Arbeit wird es dabei vorrangig um die Vermittlung von Lernzielen aus der kognitiven Domäne im Allgemeinen und die Vermittlung von prozeduralem Wissen im Speziellen gehen.

Das Medizinstudium in Deutschland gilt nach wie vor als stark verschult und theorielastig. So fordern die Medizinstudierenden im Hartmannbund und die Bundesvertretung der Medizinstudierenden in Deutschland e. V. (bvmd) in einem offenen Brief zum „Masterplan Medizinstudium 2020“ mehr Kleingruppenunterricht, praxisbezogene und interaktive Lehre sowie mehr fallorientiertes Arbeiten und problemorientiertes Lernen (Offener Brief „Masterplan Medizinstudium 2020“ 2016).

Spätestens seit Einführung des Nationalen Kompetenzbasierten Lernzielkatalogs Medizin (NKLM) im Jahr 2015 rückt neben der Lehre von faktischem Wissen auch die Vermittlung von Kompetenz zunehmend als Lernziel in den Fokus an den medizinischen Fakultäten in Deutschland. Der NKLM unterscheidet dabei drei aufeinander aufbauende Kompetenzebenen: Faktenwissen, Handlungs- und Begründungswissen und Handlungskompetenz (Letzteres unterteilt in „Unter Anleitung selbst durchführen und demonstrieren“ und „Selbstständig

[1] Aus Gründen der einfacheren Lesbarkeit werden in dieser Arbeit die männlichen Formen von Personenbeschreibungen genannt, es sind jedoch stets in vollem Umfang alle Geschlechter mit einbezogen.

und situationsadäquat in Kenntnis der Konsequenzen durchführen“) (NKLM 2015). Diese Einteilung basiert dabei auf der von Miller (1990) vorgeschlagenen Pyramide mit den vier von der Basis aufsteigenden Kompetenzen Faktenwissen (engl.: „*knows*“), prozedurales Wissen (engl.: „*knows how*“), Handlungswissen (engl.: „*shows how*“) und Anwendung von Wissen im praktischen Alltag (engl.: „*does*“).

Es stehen bereits verschiedene Lehrformate zur Verfügung, welche die Vermittlung prozeduraler Kompetenzen zum Ziel haben. Deren Implementierung und Umsetzung ist jedoch abhängig von den verfügbaren Ressourcen und findet oft in Formaten von geringer Authentizität statt.

Laut Approbationsordnung für Ärzte ist das wesentliche Ziel der ärztlichen Ausbildung „der wissenschaftlich und praktisch in der Medizin ausgebildete Arzt“ (ÄApprO 2002). Auch sind die wichtigsten von Ärzten an sich selbst gestellten Anforderungen nach Herzig et al. (2006) Wissen, dessen Umsetzung als Handeln sowie Patientenorientierung, gelebt als Empathie.

Folglich sollte in der Ausbildung von zukünftigen Ärzten der Vermittlung von prozeduralen Kompetenzen mit Verknüpfung und Anwendung von faktischem Wissen ein größerer Stellenwert beigemessen werden. Eine der wichtigsten Grundlagen für diese prozedurale Kompetenz stellt dabei das „klinische Denken“ dar, das im Folgenden treffender mit dem englischen Begriff *Clinical Reasoning* bezeichnet und näher erläutert werden soll.

1.1 *Clinical Reasoning*

Clinical Reasoning oder *Clinical Decision-Making* beschreibt den kognitiven Prozess der klinischen Entscheidungsfindung eines Arztes im Rahmen der Diagnostik und Therapie eines Patienten. Dies beinhaltet eine fundierte Wahl diagnostischer und therapeutischer Optionen unter Berücksichtigung von Risiko und Nutzen dieser sowie das Ableiten von Arbeits- und Differentialdiagnosen anhand der vom Patienten präsentierten Symptome und der gesammelten Informationen.

Die kognitiven Prozesse, welche dem *Clinical Reasoning* zugrunde liegen, werden bisher in der Literatur als ein aus zwei Komponenten bestehendes System

beschrieben; es besteht demnach aus einer intuitiven sowie einer analytischen Komponente (Dawson 1993; Hogarth 2001; Stanovich 2004; Croskerry 2009; Kassirer 2010).

Die intuitive Komponente ist gekennzeichnet durch die parallele Verarbeitung erster Eindrücke, das Wiedererkennen von Mustern und Reaktion auf Informationen, der keine bewusste Reflexion vorausgegangen ist (Stanovich 2004; Kassirer 2010). Sie arbeitet schnell und weitestgehend unterbewusst, ist jedoch anfällig für Fehler und Beeinflussung beispielsweise durch situative Umstände (Stress, Schlafmangel), Affekt oder Emotionen (Hogarth 2001; Vohs et al. 2007; Croskerry 2009).

Die analytische Komponente hingegen ist ein aktiver, bewusster, linearer Denkprozess, der deutlich langsamer als die intuitive Komponente abläuft, dafür aber robuster ist. Er kommt zum Einsatz, wenn die vom Patienten präsentierten Symptome und Informationen nicht von der intuitiven Komponente erkannt werden konnten (Croskerry 2009). Die analytische Komponente basiert auf Fachwissen, logischem Schlussfolgern, Wahrscheinlichkeiten und Entscheidungsfindung (Kassirer 2010).

Auch wenn diese beiden Komponenten in der Literatur oft getrennt beschrieben werden, gibt es wechselseitige Beziehungen zwischen ihnen (Kassirer 2010). Diese Reihenfolge und Abläufe unterscheiden sich je nach Ausbildungsstand und präsentiertem Problem. So beginnen Studierende oder Ärzte am Berufsanfang *Clinical-Reasoning*-Prozesse eher mit der analytischen Komponente, da sie die entsprechenden Muster noch nicht kennen. Erst mit zunehmender Berufserfahrung beginnen sie *Clinical-Reasoning*-Prozesse mit der intuitiven Komponente, auf die oft eine zweite, spätere Interpretation durch die analytische Komponente folgt. Das erste Ergebnis der intuitiven Komponente kann von dem Ergebnis der analytischen Komponente modifiziert oder sogar verworfen werden und so eine neue Einschätzung der Situation erzwingen. So ist eine Kontrolle der intuitiven Komponente durch die analytische Komponente gegeben (Croskerry 2009). Ebenfalls können oft mit der analytischen Komponente wiederholte klinische Entscheidungsfindungen automatisiert werden und so in die intuitive Komponente übergehen, also gelernt werden (Hogarth 2001; Stanovich 2004). Dies kann sich beispielsweise durch das Verinnerlichen von

Leitlinien zeigen. Überdies unterliegen *Clinical-Reasoning*-Prozesse nach Eva (2005) einer Kontextspezifität bezüglich situationsbedingter und persönlicher Faktoren; diese beinhalten beispielsweise die klinische Umgebung oder kürzlich erlebte Fälle sowie die jeweilige Erfahrung des Mediziners beziehungsweise dessen gegenwärtige Gedanken und Ansichten (Eva 2005).

Um *Clinical Reasoning* als eine der Schlüsselkompetenzen von werdenden Ärzten im Medizinstudium zu lehren, müssen also geeignete Lehrformen zur Vermittlung dieser Kompetenzen gefunden werden.

1.2 Lehrformen für *Clinical Reasoning*

Im Studium der Humanmedizin an den medizinischen Fakultäten Deutschlands kommen verschieden Lehrformen zum Einsatz. Dazu gehören unter anderem Vorlesungen, Seminare, Kleingruppenunterricht, Praktika, „Unterricht am Krankenbett“ (UaK), Famulaturen, Blockpraktika und das Praktische Jahr. Nicht alle Lehrformen eignen sich in gleichem Maße für die Vermittlung von *Clinical-Reasoning*-Kompetenzen. Vorlesungen können von einer Vielzahl von Studierenden besucht werden und dienen überwiegend der Strukturierung und Vermittlung von Faktenwissen. Eine aktive Partizipation der Studierenden findet – nicht zuletzt aufgrund des Frontalunterricht-Charakters – kaum statt. In Praktika liegt der Fokus auf der Vermittlung manuell-praktischer Fertigkeiten. Famulaturen, Blockpraktika und das Praktische Jahr ermöglichen sowohl den praxisnahen Erwerb ärztlicher Fertigkeiten als auch das Training von *Clinical Reasoning*, entweder durch Beobachtung von Ärzten bei der klinischen Entscheidungsfindung oder eigene Durchführung unter Supervision eines Arztes. Limitiert wird dies jedoch durch die notwendige, personalintensive Betreuung der Studierenden sowie die Umgebung mit tatsächlichen Patienten, in der nicht alle relevanten Aufgaben von Studierenden durchgeführt werden dürfen.

Die Lehrformen Seminar, Kleingruppenunterricht sowie UaK eignen sich eher, *Clinical-Reasoning*-Kompetenzen zu vermitteln. Bei einer geringen Gruppengröße kann ein reger Dialog zwischen Lehrendem und Lernenden stattfinden sowie *Clinical Reasoning* mit seiner intuitiven und analytischen Komponente an praktischen Beispielen in kontrollierter Umgebung trainiert werden.

Im Speziellen hat sich zur Lehre und zum Training von Denkprozessen, die dem *Clinical Reasoning* zuzuordnen sind, unter anderem das Format „Problemorientiertes Lernen“ (POL) etabliert. Die verschiedenen Terminologien wie Problemorientiertes Lernen, Problembasiertes Lernen (PBL, engl.: „*problem-based learning*“) oder Fallbasiertes Lernen (CBL, engl.: „*case-based learning*“) werden uneinheitlich verwendet und können nicht strikt voneinander getrennt werden (Reusser 2005). Im Folgenden wird hier jedoch nur die Bezeichnung POL verwendet. Seit seiner Konzeption Mitte der 1960er Jahre erhält POL zunehmenden Einzug in die Curricula medizinischer Fakultäten (Norman und Schmidt 1992). Norman und Schmidt fassten 1992 POL in der medizinischen Lehre wie folgt zusammen: Eine kleine Gruppe von fünf bis neun Studierenden (Barrows 1996) widmen sich unter Moderation eines Tutors einer medizinischen Patientengeschichte. Der Tutor ist in der Regel ein Mitarbeiter der Fakultät und sollte Experte sowohl für das Fachgebiet an sich als auch für die Lehrform POL sein; das heißt, dass der Tutor die Studierenden in der Bearbeitung der Fallbeispiele zwar anleitet und die Gruppe moderiert, seine fachliche Expertise jedoch nicht für eine aktive Wissensvermittlung nutzt (Dolmans et al. 2002). Die Lernenden bekommen dann eine Beschreibung der aktuellen Beschwerden und wichtiger Symptome sowie weitere relevante Informationen wie Alter, Geschlecht und Vorerkrankungen eines Patienten und sollen sich dann mit dem präsentierten klinischen Problem beziehungsweise der Fragestellung befassen. Dabei sollte das Problem so gewählt sein, dass die Lernenden es nicht sofort mit ihrem bisherigen Wissen lösen können. In der Gruppe wird das Problem dann erörtert, diskutiert und es werden Lösungsansätze entworfen (Norman und Schmidt 1992). Neben der Vermittlung von Fachwissen ist das Erlernen von *Clinical-Reasoning*-Kompetenzen ein vorrangiges Ziel von POL. Weiterhin soll auch das selbstgesteuerte Lernen gefördert werden (Norman und Schmidt 1992).

An der Universitätsmedizin Göttingen (UMG) wird POL in manchen Seminaren, Kleingruppenunterrichten und UaKs der verschiedenen Module mehr oder weniger stringent praktiziert. Als explizite Lehrform für POL existiert für die Studierenden des sechsten klinischen Semesters die Möglichkeit „Problemorientiertes Lernen, elektronische Medien assistiert“ (POLEMA) zu belegen. Dort werden Studierende aus höheren Semestern oder aus dem Praktischen Jahr als Tuto-

ren in den Kleingruppen eingesetzt, ergänzt durch regelmäßige von Experten geleitete Plenarveranstaltungen.

Kleingruppenunterricht ist personal- und damit kostenintensiv, ebenfalls ist die Qualität der Lehre stets abhängig von jedem einzelnen Lehrenden. Somit sollten weitere wissenschaftlich fundierte Methoden zur Lehre von *Clinical-Reasoning*-Kompetenzen gefunden und in Erwägung gezogen werden.

1.3 Prüfungen im Medizinstudium

Prüfungen sind ein elementarer Bestandteil des Medizinstudiums und dienen zumeist als Lernerfolgskontrolle. Neben schriftlichen Prüfungen gibt es auch praktische oder mündliche Prüfungen, jedoch soll im Folgenden nur auf Erstere eingegangen werden. Im Zuge der fortschreitenden Digitalisierung werden Prüfungen immer häufiger computerbasiert durchgeführt, statt wie bisher Papier zu verwenden. Prüfungen unterscheiden sich anhand der verwendeten Fragetypen und Prüfungskonsequenzen in ihrem Sinn und Zweck.

1.3.1 Prüfungskonsequenzen

Es kann zwischen summativen und formativen Prüfungen unterschieden werden. Summative Prüfungen ziehen eine Bewertung nach sich, erzeugen (Leistungs-) Punkte oder Noten für den Prüfling und entscheiden so gegebenenfalls über Bestehen oder Nichtbestehen. Sie haben also Konsequenzen, die von einer schlechten Bewertung über die Wiederholung einer Prüfung bis hin zum Ausscheiden aus dem Studiengang reichen können. Formative Prüfungen hingegen sind unbenotet. Sie generieren für den Prüfling Feedback über seinen individuellen Wissensstand, können ihm so eigene Schwächen aufzeigen und die Grundlage für weitere Lernaktivitäten bilden. Von Raupach et al. (2013) konnte gezeigt werden, dass die Art einer Prüfung, ob summativ oder formativ, maßgeblich das Lernverhalten der Studierenden und damit das Ergebnis einer Prüfung beeinflusst. In diesem Zusammenhang wurde das Axiom *„assessment drives learning“* geprägt. Summative Prüfungen generieren einen größeren Lernanreiz und -erfolg als formative Prüfungen, unabhängig von der Lehre (Raupach et al. 2013). Zunächst sollen die verschiedenen Item- und Fragetypen erläutert werden.

1.3.2 Items und Fragetypen

Ein Item besteht in der Regel aus einem Fragestamm, der eine einleitende sogenannte Vignette mit kurzer Beschreibung eines Fallbeispiels sowie die eigentliche Frage enthält. Je nach Fragetyp folgt dann eine Liste von Antwortoptionen. Allgemein kann dabei zwischen richtigen Antworten und Distraktoren unterschieden werden. Funktionale Distraktoren (engl.: „*functional-distractors*“) sind Falschantworten, die von der richtigen Antwort ablenken sollen, ohne jedoch völlig unplausibel oder trivial zu sein. Trifft letzteres zu, entspräche dies einem nicht-funktionalen Distraktor (engl.: „*non-functional-distractor*“), der oft auch von leistungsschwachen Prüflingen sofort ausgeschlossen werden kann und somit die Trennschärfe eines Items verringert. (Krebs 2004; Brüstle 2011)

Um den sogenannten *Cueing*-Effekt zu vermeiden muss bei der Formulierung von Fragestamm und Antworten darauf geachtet werden, dass diese keine Hinweise auf die Lösung desselben oder eines anderen Items enthalten.

Im Studium der Humanmedizin an der UMG werden wie in den anderen medizinischen Fakultäten Deutschlands in den schriftlichen Prüfungen überwiegend Einfachauswahlfragen verwendet. Auch die schriftlichen Teile des ersten sowie zweiten Abschnitts der ärztlichen Prüfung bestehen ausschließlich aus Fragen dieses Typs (Brüstle 2011). Oft wird dieser klassische Fragentyp, aus dem Englischen übernommen, fälschlicherweise als *Multiple Choice* bezeichnet, obgleich die korrekte Bezeichnung Einfachauswahlfrage oder *Single Choice* wäre. Der folgende Abschnitt zeigt eine Übersicht über die wichtigsten im Medizinstudium eingesetzten Fragetypen.

Bei einer **Einfachauswahl- oder „Typ-A“-Frage** (engl.: „*multiple choice question*“) muss der Prüfling die eine bestmögliche Antwort aus einer Liste von Antwortmöglichkeiten wählen (Brüstle 2011). In der Regel werden fünf Antwortmöglichkeiten vorgegeben, jedoch kann auch mit nur vier oder drei Antwortmöglichkeiten gearbeitet werden (Rodriguez 2005; Rogausch et al. 2010). Je nach Formulierung der Frage soll die am ehesten richtige („Typ A_{pos}“) oder am ehesten falsche („Typ A_{neg}“) Aussage vom Prüfling gewählt werden. Letzterer Typ sollte jedoch nur mit Bedacht und nicht zu häufig in einer Prüfung einge-

setzt werden, zum Beispiel in Fällen, in denen das Kennen einer wichtigen Ausnahme entscheidend ist (Krebs 2004).

Der Unterschied gegenüber einer Einfachauswahlfrage besteht bei einer **Mehrfachauswahl- oder „Pick-N"-Frage** (engl.: „*multiple response question*") darin, dass der Prüfling mehrere bestmögliche Antworten (in der Regel zwei bis fünf) aus einer Liste von Antworten wählen soll. Die erwartete Zahl von Antworten wird in der Regel angegeben. (Krebs 2004; Bauer et al. 2011)

Bei einer ***Multiple-True-False*- (MTF) oder „K-prim"-Frage** muss der Prüfling in einem Item bei jeder von mehreren gegebenen Antwortmöglichkeiten entscheiden, ob diese richtig oder falsch ist. (Weih et al. 2009)

Bei einer **Kurzantwortfrage** (engl.: „*short answer question*" oder „*short essay question*") soll der Prüfling die richtige Antwort als Freitext formulieren. Die erwartete Antwortlänge variiert je nach Fragestellung von einzelnen Wörtern bis hin zu wenigen Sätzen. Dieser Fragetyp findet überwiegend im angelsächsischen Raum Verwendung (Schulze und Drolshagen 2006).

Analog zur Kurzantwortfrage soll der Prüfling bei einer **Essay-Frage** ebenfalls die richtige Antwort als Freitext formulieren. Im Fokus steht hier jedoch „die schriftliche Darlegung von Zusammenhängen oder die Begründung einer Entscheidung" (Schulze und Drolshagen 2006). Die erwartete Antwortlänge bei diesem Fragetyp fällt daher länger aus.

Das sogenannte ***Long Menu*** ist ein Fragetyp, welcher in computerbasierten Prüfungen Verwendung findet. Im Computersystem ist eine alphabetisch sortierte Antwortliste mit Begriffen hinterlegt, welche alle richtigen Antworten, funktionale, aber auch nichtfunktionale Distraktoren und entsprechende Synonyme beinhaltet. Diese Liste sollte mindestens 500 Begriffe enthalten um die Wahrscheinlichkeit eines *Cueing*-Effekts zu verringern (Schuwirth et al. 1996; Kopp et al. 2006). Der Prüfling sieht zunächst ein leeres Texteingabe-Feld. Gibt der Prüfling mindestens drei Buchstaben der von ihm gewünschten Antwort ein, erscheinen alle Einträge der Antwortliste, die diese Buchstaben in dieser Reihenfolge enthalten. Er kann nun einen der hinterlegten Begriffe als Antwort auswählen. Sollte eine der Meinung des Prüflings nach richtige Antwort nicht in der Antwortliste enthalten sein, gibt es die Möglichkeit, in einem weiteren Frei-

textfeld die vom Prüfling gewünschte Antwort einzugeben, da im *Long Menu* nur bereits hinterlegte Antworten gewählt werden können. Es können auch mehrere Begriffe als korrekte Antwort definiert werden, von denen der Prüfling nur einen korrekt nennen muss.

Der Vorteil von *Long-Menu*-Fragen ist die einfache computergestützte Auswertung sowie das Fehlen eines *Cueing*-Effekts (Schuwirth et al. 1996).

1.3.3 Prüfung aktiven und passiven Wissens

Bei den verschiedenen Fragetypen lassen sich sogenannte *Recognition Tests* von *Production Tests* unterscheiden. Bei den *Recognition Tests* muss der Prüfling passiv die korrekte Antwort aus den gegebenen Antwortmöglichkeiten wiedererkennen und auswählen. Dazu zählen Einfach- sowie Mehrfachauswahl- oder K-prim-Fragen. Bei den sogenannten *Production Tests* hingegen muss die korrekte Antwort vom Prüfling aktiv produziert werden. Hierzu zählen Kurzantworten, Essays oder *Long-Menu*-Fragen (Larsen et al. 2008).

Recognition Tests eignen sich zur Prüfung von Faktenwissen auf niedrigem taxonomischen Niveau nach Bloom, wohingegen komplexere kognitive Fähigkeiten wie *Clinical Reasoning*, also prozedurales Wissen, eher durch *Production Tests* wie *Key-Feature*-Fälle mit *Long-Menu*-Fragen geprüft werden können (Kopp et al. 2006).

1.4 *Test-enhanced Learning*

Aus der Forschung ist bekannt, dass Lerninhalte, die wiederholt geprüft werden, besser behalten werden als solche, die wiederholt gelernt werden. Diesem Phänomen liegt der sogenannte *Testing Effect* zugrunde. Roediger und Karpicke verfassten hierzu 2006 eine umfangreiche Übersichtsarbeit mit dem Titel „*The Power of Testing Memory*“ (Roediger und Karpicke 2006).

Es können dabei indirekte Effekte von direkten Effekten durch Testen unterschieden werden. Die indirekten Effekte beziehen sich dabei auf Prozesse neben dem eigentlichen Vorgang des Geprüftwerdens. Dazu zählt unter anderem der Wissenszuwachs aus dem in den Prüfungen gegebenen Feedback, die kontinuierliche, wiederholte Beschäftigung mit einem Thema, statt kurzfristigen

Lernens vor einer Klausur oder die bessere Orientierung im Lernstoff durch individuelle Rückmeldung, welches Wissen man schon beherrscht und welches noch nicht (Roediger und Karpicke 2006).

Im Gegensatz dazu steht der direkte *Testing Effect*, der durch den Abruf des Lernstoffs während einer Prüfung auftritt. Roediger und Karpicke fassen viele der von ihnen untersuchten Studien wie folgt zusammen: Studierende einer Gruppe „Prüfen“ lernten einen bestimmten Lernstoff und wurden dann einer initialen Prüfung (in manchen Studien auch mehreren Prüfungen) unterzogen. Als primärer Ausgangspunkt wurde die Retention des Gelernten in einem Abschlusstestat gemessen und mit der Retention von verschiedenen Kontrollgruppen verglichen. Während eine Art von Kontrollgruppe („keine Prüfung“) den Lernstoff lediglich einmal lernte und das Abschlusstestat schrieb, ohne eine initiale Prüfung zu bearbeiten, lernte eine zweite Art von Kontrollgruppe („wiederholt Lernen“) den Lernstoff ein zweites Mal, bevor sie das Abschlusstestat schrieb. Für die zweite Art von Kontrollgruppe wurde so sichergestellt, dass die Expositionszeit mit dem Lernstoff zwischen den Gruppen „Prüfen“ und „wiederholt Lernen“ gleich war. Das typische Ergebnis in der Literatur ist, dass die Leistung in der Gruppe „Prüfen“ die Leistungen in beiden Kontrollgruppen übertrifft (Roediger und Karpicke 2006).

Von Wing et al. (2013) konnte der *Testing Effect* auch mittels funktioneller Magnetresonanztomographie nachvollzogen werden. Es zeigte sich eine gesteigerte Aktivität im Hippocampus, dem Temporallappen und dem präfrontalen Kortex für in einem finalen Test richtig erinnerte Wörter, welche nach initialem Lernen in einem zweiten Schritt geprüft wurden, gegenüber Wörtern, die erneut gelernt wurden (Wing et al. 2013).

Theoretisch könnte der *Testing Effect* auch nicht wie beschrieben durch das Abrufen des Lernstoffs selbst (*Retrieval Hypothesis*), sondern durch eine zusätzliche Expositionszeit gegenüber dem Lernstoff (*Total-Time Hypothesis*) zu erklären sein. Jedoch konnte letzteres in Studien, in denen die Expositionszeit kontrolliert wurde, nicht bestätigt werden (Schmidmaier et al. 2011; Dobson und Linderholm 2015). Auch in der Vorstudie zu dieser Arbeit war der nachgewiesene Effekt robust gegenüber der Adjustierung für die Bearbeitungszeit (Raupach et al. 2016).

In Studien, in denen die Anzahl der Wiederholungen der Prüfungen variiert wurde, konnte gezeigt werden, dass eine größere Anzahl von Wiederholungen zu einer höheren langfristigen Retention führt (Karpicke und Roediger 2007). Ebenfalls korrelierte in einer Studie von Larsen et al. die Anzahl korrekter Abfragen eines Items mit der erfolgreichen Abfrage desselben Items in einer finalen Prüfung (Larsen et al. 2013b). Um also einen langfristigen Lernerfolg sicherzustellen reicht ein einmaliges Prüfen nicht aus. Ebenfalls gibt es Hinweise in der Literatur darauf, dass auch der zeitliche Abstand zwischen den Prüfungen bzw. Abfragen einen Einfluss auf den Lernerfolg hat; so führt ein zeitlicher Abstand zu höherer Retention, vorausgesetzt die Zeit zwischen initialem Lernen und erster Abfrage ist nicht so lang, dass der Inhalt wieder vergessen wurde (Roediger und Karpicke 2006). Ebenfalls gibt es nach der „*retrieval effort hypothesis*“ Hinweise darauf, dass Inhalte, die im Rahmen von *Test-enhanced Learning* gelernt werden, besser behalten werden, wenn die initiale Abfrage schwierig aber erfolgreich ist (Carpenter 2009; Pyc und Rawson 2009).

Während die meisten Studien zum *Testing Effect* mit wenig komplexem Lernstoff von niedrigem taxonomischen Niveau wie dem Lernen von Wortlisten oder Wortverknüpfungen durchgeführt wurden, konnte der *Testing Effect* auch außerhalb von „Laborbedingungen“ in anwendungsnahen Bereichen gezeigt werden. So beispielsweise bei Dobson und Linderholm in einem Anatomie- und Physiologie-Kurs, in welchem gelesene Lerninhalte, die unmittelbar geprüft wurden, signifikant besser behalten wurden als solche, die stattdessen wiederholt gelesen wurden, oder solche, zu denen die Studierenden sich Notizen machen sollten (Dobson und Linderholm 2015). Kromann et al. konnten den *Testing Effect* auch für praktische Fertigkeiten zeigen. In ihrer Studie erzielten Studierende eines siebten Semesters zwei Wochen nach Intervention signifikant bessere Ergebnisse in kardiopulmonaler Reanimation, wenn diese nach Unterricht und Training (3,5 Stunden) eine formative Prüfung (0,5 Stunden) ablegen mussten, im Vergleich zu Studierenden, die lediglich Unterricht und Training (vier Stunden) ohne abschließende Prüfung erhielten. Aus den Zeitangaben wird ersichtlich, dass auch hier beide Gruppen dieselbe Expositionszeit gegenüber dem Lernstoff hatten (Kromann et al. 2009). In einer Folgestudie zeigten Kromann et al., dass sich dieser Effekt mit einer Effektstärke von 0,4 auch noch

nach einem halben Jahr nachweisen ließ, jedoch war der Unterschied zwischen den Gruppen nicht mehr statistisch signifikant, was am ehesten mit der zu geringen Power dieser Analyse erklärbar war (Kromann et al. 2010).

Die Studie von Kromann et al. 2010 stellt hierbei eine Ausnahme dar, bei der die Retention auch nach einem halben Jahr untersucht wurde. Bei der Mehrzahl der Studien zum *Testing Effect* wurde die Retention meist bereits nach sehr kurzer Zeit (Tage bis Wochen, z. B. sieben Tage (Logan et al. 2011; Baghdady et al. 2014)) untersucht.

Ebenfalls finden in den meisten Studien zum *Testing Effect* als Intervention oder finale Prüfung reproduktive Prüfungsformate von niedrigem taxonomischen Niveau wie Einfachauswahlfragen (Baghdady et al. 2014) Anwendung. Sowohl Roediger und Karpicke als auch Larsen et al. fassten jedoch zusammen, dass *Production-Tests* (vgl. Kapitel 1.3.3) einen größeren Effekt auf die Retention haben als *Recognition-Tests* (Roediger und Karpicke 2006; Larsen et al. 2008). Untersucht wurde dies unter anderem in einer Studie von Butler und Roediger, in welcher die Retention der Inhalte aus drei Kunstgeschichte-Vorlesungen gemessen wurde. Beantworteten Probanden unmittelbar nach einer Vorlesung Fragen im Kurzantwortformat, also einem *Production-Test*, war einen Monat später in einer erneuten Prüfung die Retention signifikant größer als nach der Beantwortung von Mehrfachauswahlfragen, also einem *Recognition-Test,* oder dem Lesen einer Zusammenfassung der Vorlesung. Letztere unterschieden sich nicht in der erzeugten Retention, waren aber im Vergleich zu keiner Aktivität nach einer Vorlesung immer noch überlegen (Butler und Roediger 2007).

Prüfungen können folglich nicht nur summativ zur Bewertung von Studierenden eingesetzt werden, sondern ebenfalls formativ, als Lehrinstrument, um eine größere Retention des Lernstoffs zu begünstigen (Larsen et al. 2008). Zusammenfassend sollte *Test-enhanced Learning* also ein produktives Format annehmen sowie wiederholt und mit ausreichenden zeitlichen Zwischenräumen durchgeführt werden. Hierfür eignet sich beispielsweise das *Key-Feature*-Format.

1.5 *Key Features*

In Kanada wurden bis 1992 im *Canadian Qualifying Examination in Medicine*, welches alle Absolventen medizinischer Hochschulen durchlaufen müssen, sogenannte *Patient Management Problems* (PMPs) als Prüfungsformat für *Clinical-Reasoning*-Kompetenzen verwendet (Page et al. 1995). Diese wurden von McGuire entwickelt und bestehen in der Regel aus umfassenden klinischen Problemen mit einer Fallbeschreibung sowie entsprechenden Fragen zu Anamnese, Untersuchung und Diagnose (McGuire et al. 1976 zitiert nach Page und Bordage 1995). Ein einzelnes PMP zu bearbeiten konnte bis zu 90 Minuten dauern (McGuire et al. 1976 zitiert nach Farmer und Page 2005); insgesamt waren zwischen 12 und 15 PMPs in ca. dreieinhalb Stunden zu bearbeiten (Page und Bordage 1995). Studien zeigten jedoch, dass diese PMPs ungeeignet waren, *Clinical-Reasoning*-Kompetenzen zu prüfen, da es sich hierbei entgegen früherer Meinung nicht um eine allgemeine Fähigkeit handelt und Inhaltsspezifität ein wichtiger Faktor für *Clinical Reasoning* ist; außerdem verfügten PMPs nur über eine geringe Reliabilität (Page und Bordage 1995; Farmer und Page 2005). Auf der Suche nach einem alternativen Prüfungsformat für *Clinical-Reasoning*-Kompetenzen im Rahmen eines sechsjährigen Forschungsprojekts 1986–1992 für das *Medical Council of Canada* prägten Bordage und Page den Begriff *Key Feature* (Bordage und Page 1987 zitiert nach Farmer und Page (2005); Page und Bordage 1995). Die zu lösenden klinischen Probleme können und sollten auf die zur Lösung des Problems notwendigen Schlüsselstellen oder eben „*Key Features“* reduziert werden: „*A key feature is defined as a critical step in the resolution of a problem*“ (Page et al. 1995). Zwei Zusätze ergänzen die Definition eines *Key Features*: „*(1) it focuses on a step in which examinees are most likely to make errors in the resolution of the problem, and (2) it is a difficult aspect of the identification and management of the problem in practice*“ (Page et al. 1995). Seit 1992 enthalten die Prüfungen des *Medical Council of Canada* auch *Key-Feature*-Fälle (Hatala und Norman 2002).

Dabei sind *Key Features* mehr als generelles Konzept sowohl für Prüfungen als auch für die Lehre und nicht nur als Prüfungsformat an sich zu verstehen; theoretisch ließen sich auch Stationen einer *Objective Structured Clinical Examination* (OSCE) um die *Key Features* eines klinischen Szenarios erstellen

(Hrynchak et al. 2014). Im Allgemeinen versteht man unter einem *Key Feature* jedoch einen Teil einer zumeist schriftlichen Prüfung, welcher eine klinische Fallbeschreibung und eine oder mehrere darauffolgende Fragen zu der Fallbeschreibung beinhaltet. Die Fallbeschreibung oder sogenannte Fallvignette enthält dabei in der Regel relevante, bisweilen auch weniger relevante Informationen, wie Anamnese, Symptome und klinische Befunde. Diese sollten kurz und uninterpretiert (z. B. Herzfrequenz 160/Minute statt „Tachykardie“) sein (Schuwirth und van der Vleuten 2004). Die darauffolgenden einzelnen *Key-Feature*-Fragen sollten auf die Schlüsselstellen und wichtigen Entscheidungen der klinischen Fallgeschichte ausgerichtet sein und kein Faktenwissen prüfen. Diese sollten die Prüflinge dann mit den gegebenen Informationen im Kontext der Fallgeschichte beantworten und eine klinische Entscheidung treffen (Hrynchak et al. 2014). In der vorliegenden Arbeit soll folgende Terminologie verwendet werden: Eine *Key-Feature*-Prüfung besteht aus mehreren *Key-Feature*-Fällen. Jeweils ein *Key-Feature*-Fall beschreibt ein medizinisches Fallbeispiel und enthält mehrere *Key-Feature*-Items und somit *Key-Feature*-Fragen. Letztere beiden Begriffe werden bisweilen synonym verwendet.

In einer Übersichtsarbeit mit 21 eingeschlossenen Artikeln zum Einsatz von *Key Features* im Medizinstudium schlussfolgern Hrynchak et al., dass *Key-Feature*-Fälle dazu geeignet sind, *Clinical-Reasoning*-Kompetenzen zu prüfen (Hrynchak et al. 2014).

An der UMG wurden computergestützte *Key-Feature*-Fälle erstmals im Wintersemester 2007/08 im Rahmen eines Forschungsprojektes eingesetzt (Raupach et al. 2009). Im Sommersemester 2013 folgte die Implementierung und Pilotierung von *Key-Feature*-Fällen in ihrer aktuellen Form in die Lehr-Informationstechnik der UMG. Im Wintersemester 2013/14 schloss sich eine erste prospektive, randomisierte Cross-over-Studie zum Beitrag des direkten *Testing Effects* zur mittelfristigen Retention differentialdiagnostischer und -therapeutischer Kompetenzen an (Raupach et al. 2016). In dieser Studie wurden den Studierenden in zehn elektronischen Fallseminaren (E-Fallseminaren) im wöchentlichen Wechsel medizinische Fallbeispiele aus der Inneren Medizin im *Key-Feature*-Format mit intermittierenden *Long-Menu*-Fragen („Fragefall“) oder zum Lesen („Lesefall“) präsentiert. Die gegebenen

Informationen waren für beide Präsentationsformen identisch, jedoch mussten bei den Fragefällen an entsprechenden Schlüsselstellen der Fallbeispiele Fragen beantwortet werden, während bei den Lesefällen die Informationen lediglich von den Studierenden gelesen wurden. Die Retention wurde in Abhängigkeit davon untersucht, ob ein Item zuvor im Rahmen eines Fragefalls bearbeitet, also geprüft, wurde oder als Lesefall gelesen wurde. Es zeigte sich sowohl zum Ausgangstest, unmittelbar nach dem Ende der Vorlesungszeit des Semesters, als auch zum Retentionstest nach sechs Monaten eine signifikant bessere studentische Leistung für Inhalte, die mehrfach geprüft worden waren, verglichen mit Inhalten, die mehrfach gelesen wurden. Es handelte sich dabei um die erste Studie, die den *Testing Effect* auch für *Clinical-Reasoning*-Kompetenzen unter der Benutzung von *Key-Feature*-Fällen zeigen konnte.

1.6 Videos in der Lehre

Während sich Texte zwar mit wenig Aufwand auf relevante Informationen durchsuchen lassen, ist es möglich, dass es den Studierenden dabei schwerer fällt, sich auf das klinische Fallbeispiel einzulassen und sich darin zu vertiefen. In einer psychologischen Studie konnte gezeigt werden, wie wichtig die Ähnlichkeit zwischen Lernumgebung und Situation ist, in der das Gelernte wieder abgerufen wird (Godden und Baddeley 1975). Nach der „Dualen Kodierungstheorie“ (engl.: „*dual-coding theory*“) von Paivio (1971) werden multimediale Inhalte kognitiv über zwei verschiedene Systeme verarbeitet: das verbale System für Sprache und akustische Reize sowie das visuelle System für Bilder und Schrift. Die Informationen der beiden Systeme werden dann im Gehirn miteinander integriert und weiterverarbeitet. Auf dieser Theorie aufbauend konnten weiterführende Studien zeigen, dass simultan präsentierte audiovisuelle Inhalte eine höhere Retention von gelernten Inhalten erzeugen als verbale Inhalte allein oder eine konsekutive Präsentation von visuellen und verbalen Inhalten (Mayer und Sims 1994).

Um die Authentizität von medizinischen Fallbeispielen in der Lehre gegenüber dem Textformat zu erhöhen bietet sich das Videoformat an. Im Zuge der fortschreitenden Digitalisierung wird die Produktion, Verbreitung und Wiedergabe von Videos zunehmend einfacher. Videos sind auch bei größeren Kohorten so-

wie wiederholt einsetzbar und dabei weniger personal- und ressourcenintensiv als das Training derselben Anzahl von Studierenden unter Einsatz von Schauspielpatienten zu trainieren. Bei Praktika auf Station, die ebenfalls personalintensiv sind, kann nicht sichergestellt werden, dass zum Zeitpunkt des Praktikums auch Patienten mit denjenigen Erkrankungen, die nach Curriculum zu lehren sind, zur Verfügung stehen.

Auch sollte die Präferenz der Zielgruppe, das sind die Studierenden, nicht außer Acht gelassen werden. Hierzu wurden Studierende und Lehrende an der medizinischen Fakultät der Universität Hongkong nach deren Präferenz des Präsentationsformats der Fallbeispiele im Rahmen von POL-Lehrveranstaltungen befragt. Mehr als 75 % der Teilnehmer bevorzugten hier das Videoformat gegenüber dem Textformat (Chan et al. 2010). Auch in einer weiteren Studie aus den USA bevorzugten Studierende und Lehrende videobasierte Fallbeispiele gegenüber textbasierten (Basu Roy und McMahon 2012).

Es gibt jedoch Hinweise, dass eine höhere Authentizität auch mit einer höheren kognitiven Belastung einhergehen kann. Nach der Theorie der kognitiven Belastung (engl.: „*cognitive load theory*“) von Sweller (1998) ist die Kapazität des Arbeitsgedächtnisses begrenzt. Mousavi et al. (1995) konnten zwar zeigen, dass die kognitive Belastung bei gleichzeitiger Präsentation von kongruenten verbalen und visuellen Inhalten reduziert wird, jedoch wurden Audioaufnahmen, die zusammen mit Bildern präsentiert wurden, untersucht, nicht Videos. Wird die Kapazitätsgrenze nach der Theorie der kognitiven Belastung überstiegen, könnte dies andererseits einem potentiell positiven Effekt einer gesteigerten Authentizität auf die Retention von gelernten Inhalten entgegenwirken (La Rochelle et al. 2011).

La Rochelle et al. (2011; 2012) untersuchten die Zusammenhänge von Authentizität verschiedener Lehrformate und deren Wirkung an Medizinstudierenden im zweiten Studienjahr. Dabei wurden das Lesen papierbasierter Fallbeschreibungen, das Anschauen von Videos gefilmter Schauspielpatient-Arzt-Interaktionen sowie das reale Erleben von Schauspielpatient-Arzt-Interaktionen untereinander verglichen. Alle drei Lehrformate enthielten Anamnese und klinische Untersuchung der Patienten und waren inhaltlich identisch. In Kleingruppen wurden die Fallgeschichten nach entsprechender Präsentation diskutiert

und besprochen. Die Leistungen der Studierenden wurden anhand von OSCEs, Video-Quiz und schriftlichen Prüfungen erhoben. Es konnte hier keine signifikante Verbesserung der studentischen Leistungen für Lehrformate mit größerer Authentizität gefunden werden (La Rochelle et al. 2011; 2012).

In der Literatur besteht jedoch keine Einigkeit darüber, ob die Präsentation von Fallbeispielen im Videoformat tiefgreifendes Denken (engl.: „*deep critical thinking*“) fördert (Kamin et al. 2003; Balslev et al. 2005) oder behindert (Basu Roy und McMahon 2012). Basu Roy und McMahon schlussfolgerten, dass die zusätzlichen visuellen Informationen von den medizinischen Inhalten ablenken. Allerdings ist ebenso vorstellbar, dass durch die Präsentation der Inhalte im realen Anwendungskontext eine tiefere Reflexion befördert wird, welches nach Koole et al. (2012) mit einer größeren Leistung beim Lösen von Fallbeispielen korreliert.

In jedem Fall spielt die Qualität der verwendeten Lehrvideos sowie deren Implementierung eine wichtige Rolle. Dong und Goh (2015) veröffentlichten zwölf Tipps zum effektiven Einsatz von Videos in der medizinischen Lehre, von denen hier exemplarisch einige wichtige Aspekte aufgegriffen werden sollen.

Tipp 1: Durch realitätsnahe Darstellung von klinischen Situationen erzeugen Videos Neugierde in den Studierenden, welche wiederum das Lernen begünstigt (Pluck und Johnson 2011).

Tipp 5: Geplante, intermittierende Pausen zwischen Videos mit aktiver Beteiligung der Studierenden, beispielsweise der Beantwortung von Fragen, helfen, eine tiefergehende Beschäftigung mit den Inhalten zu fördern. Passives Ansehen von Videos hingegen scheint nicht hinreichend effektiv zu sein, um schwierige Inhalte oder abstrakte Themen zu verstehen oder kritisches Denken zu begünstigen.

Tipp 6: Die Verwendung von Videos sollte bezüglich der Lernziele und Inhalte an das restliche Curriculum angeglichen sein (Biggs 2007 zitiert nach Dong und Goh 2015).

Tipp 8: Eine kognitive Überbeanspruchung der Studierenden, zum Beispiel durch inkohärente Informationen, sollte vermieden werden, da dies einen nach-

teiligen Effekt auf den Lernerfolg hat (Mayer und Moreno 2003 zitiert nach Dong und Goh 2015).

Tipp 11: Identifiziere glaubwürdige Videos von professioneller Qualität.

Tipp 12: Verwendete Videos sollten von hoher Audio- und Videoqualität sein. Falls diese selbst erstellt werden sollen, müssen entsprechende Fähigkeiten, Erfahrung sowie technische Ausstattung vorhanden sein. Es sollten Drehbücher geschrieben und verwendet werden. Zur Qualitätssicherung sollte selbst erstelltes Videomaterial von erfahrenen Kollegen angesehen und überprüft werden. (Dong und Goh 2015)

1.7 Ziele und Hypothesen

Vor dem Hintergrund der hier vorgestellten aktuellen Datenlage können *Key-Feature*-Fälle als ein geeignetes Format für die Prüfung von *Clinical Reasoning* angesehen werden, ebenfalls konnte der *Testing Effect* für prozedurale Kompetenzen bei wiederholtem Prüfen im *Key-Feature*-Format gezeigt werden. Auch eignen sich Videos, um die Authentizität in der Präsentation medizinischer Fallbeispiele gegenüber dem Textformat zu erhöhen.

Dem Verfasser dieser Arbeit sind jedoch keine Untersuchungen bekannt, welche die Einflüsse verschiedener Präsentationsformate, wie etwa Text oder Video, auf die kurz- und mittelfristige Retention von *Clinical-Reasoning*-Kompetenzen quantifiziert hätten. Auch ergab die Literaturrecherche keine Arbeiten, in denen die Verwendung von Video-*Key-Feature*-Fällen in der Lehre von *Clinical Reasoning* untersucht worden wäre.

Somit ergeben sich die primäre Forschungsfrage Nr. 1 und die explorativen Forschungsfragen Nrn. 2–5:

1. Wie wirkt sich das Präsentationsformat der Inhalte (Text- oder Video-*Key-Feature*) auf den Erwerb und die kurz- sowie mittelfristige Retention komplexer kognitiver Fertigkeiten (hier: Differentialdiagnostik und -therapie) aus?

2. Welchen Einfluss hat die Kongruenz zwischen den Präsentationsformaten während der Lernphase und in den konsekutiven Prüfungen auf die kurz- sowie mittelfristige Retention?

3. Inwiefern unterscheiden sich die Studierenden bezüglich des Geschlechts, des Alters und der Klausurvorleistung hinsichtlich der Effekte der Präsentationsformate auf ihren individuellen Lernerfolg?

4. Wie unterscheiden sich die beiden Präsentationsformate Text- und Video-*Key-Feature* hinsichtlich ihrer Itemkennwerte und Testgütekriterien?

5. Wie beurteilen die Studierenden die unterschiedlichen Präsentationsformate Text- und Video-*Key-Feature*?

Aus den Forschungsfragen lassen sich für diese Arbeit folgende fünf Hypothesen ableiten:

1. Inhalte, die im Rahmen von Video-*Key-Feature*-Prüfungen wiederholt wurden, werden kurz- sowie mittelfristig nicht besser behalten als Inhalte, die mit Text-*Key-Feature*-Prüfungen wiederholt wurden.

2. Die kurz- sowie mittelfristige Retention ist nicht höher, wenn die Präsentationsformate während der Lernphase und in den konsekutiven Prüfungen kongruent sind.

3. Die Studierenden unterscheiden sich bezüglich des Geschlechts, des Alters und der Klausurvorleistung nicht hinsichtlich der Effekte der Präsentationsformate auf ihren individuellen Lernerfolg.

4. Die beiden Präsentationsformate Text- und Video-*Key-Feature* unterscheiden sich nicht hinsichtlich ihrer Itemkennwerte und Testgütekriterien.

5. Die Studierenden beurteilen die unterschiedlichen Präsentationsformate Text- und Video-*Key-Feature* nicht unterschiedlich.

2 Material und Methoden

Im Rahmen der hier vorgestellten Arbeit wurden zwei prospektive, randomisierte Cross-over-Studien im Wintersemester 2014/15 und im Sommersemester 2015 an der Universitätsmedizin Göttingen (UMG) durchgeführt. Studienleiter dieser Studien war Prof. Dr. Tobias Raupach. Zur Beantwortung der Forschungsfragen kamen quantitative und qualitative Methoden zum Einsatz. Im Folgenden werden zunächst die Gemeinsamkeiten der durchgeführten Studien erläutert, in den Kapiteln 2.1.1 und 2.1.2 werden anschließend die Unterschiede dargestellt. Eine detaillierte Auflistung der verwendeten Software befindet sich in Tabelle A1 im Anhang.

2.1 Studiendesign

In beiden Cross-over-Studien wurden den Studierenden des dritten klinischen Semesters in zwölf wöchentlichen elektronischen Fallseminaren (E-Fallseminaren) medizinische Fallbeispiele im *Key-Feature*-Format präsentiert. Dabei kamen Video-*Key-Feature*-Items als Interventionsitems sowie Text-*Key-Feature*-Items als Kontrollitems zum Einsatz. Die Exposition gegenüber den beiden Formaten wurde in dieser Studie nicht auf der individuellen Ebene (Studienteilnehmer) sondern auf Item-Ebene vorgenommen. Folglich bearbeiteten alle Studierenden sowohl Interventions- als auch Kontrollitems in den E-Fallseminaren. Die Studierenden wurden zunächst nach Geschlecht und bisheriger Studienleistung stratifiziert und randomisiert in zwei Gruppen A und B eingeteilt. Die Gruppen- und damit auch die Raumeinteilung blieb über den gesamten Studienverlauf gleich. Abbildung 1 zeigt das Schema des Studiendesigns, welches im Folgenden beschrieben wird.

In der ersten Semesterwoche fand im ersten E-Fallseminar ein für die Studierenden beider Gruppen identischer formativer Eingangstest statt. Dieser bestand aus vier *Key-Feature*-Fällen im Textformat mit jeweils zweimal sechs und zweimal neun *Key-Feature*-Items, insgesamt also 30 *Key-Feature*-Items. Der Eingangstest diente dazu, den je nach individueller Vorerfahrung (beispielsweise durch Berufserfahrung oder Famulaturen) verschiedenen Wissensstand der Studierenden vor der Intervention zu untersuchen. Außerdem sollte so die Ver-

gleichbarkeit des studentischen Leistungsniveaus in den Gruppen A und B überprüft werden.

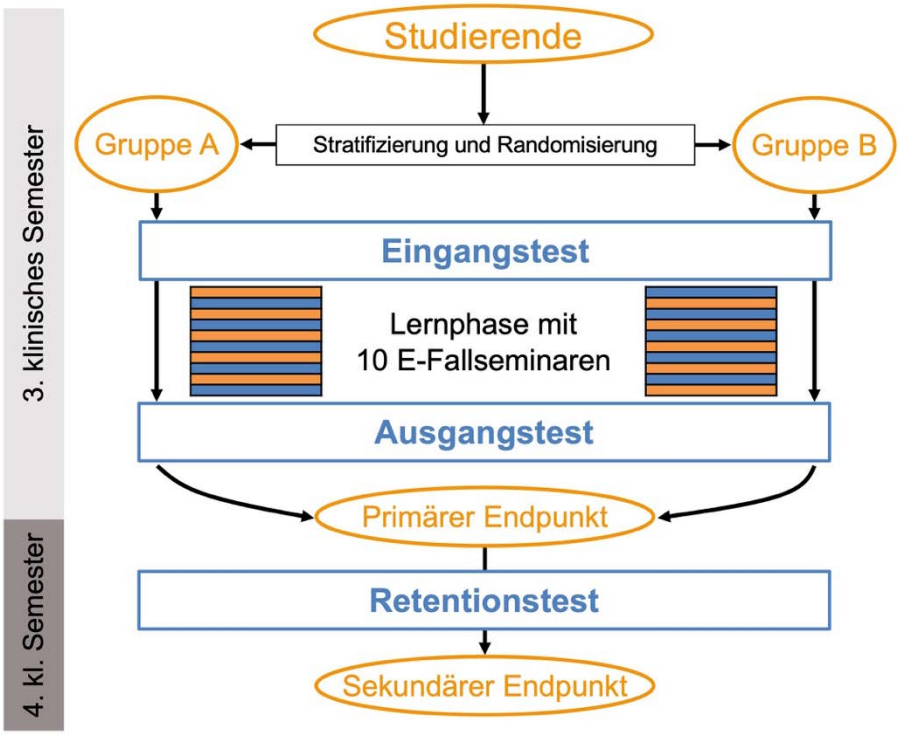


Abbildung 1: Schema des Studiendesigns

Es folgte die Lernphase in den Semesterwochen zwei bis elf. In jedem E-Fallseminar während der Lernphase bearbeitete jeweils eine Gruppe in einem der beiden Räume *Key-Feature*-Fälle im Textformat (blaue Balken in Abbildung 1), während die andere Gruppe in dem anderen Raum inhaltlich identische *Key-Feature*-Fälle im Videoformat (orangefarbene Balken in Abbildung 1) bearbeitete. Die Zuordnung des Präsentationsformats zu den Computerräumen (und somit zu den Gruppen A und B) geschah im wöchentlichen Wechsel. Die E-Fallseminare bestanden jeweils aus drei *Key-Feature*-Fällen mit je fünf *Key-Feature*-Items. Ein Student der Gruppe A, der an allen zehn E-Fallseminaren der Lernphase teilnahm, bearbeitete folglich 30 *Key-Feature*-Fälle (150 *Key-Feature*-Items), davon 15 Fälle (75 Items) im Textformat und 15 Fälle (75 Items) im Videoformat. Ein Student der Gruppe B bearbeitete die gleiche Anzahl an Items; die Exposition war hinsichtlich des Präsentationsformats jedoch jeweils komplementär zur Präsentation in Gruppe A.

Im letzten E-Fallseminar des Semesters in Semesterwoche zwölf bearbeiteten alle Studierende einen für beide Gruppen identischen formativen Ausgangstest.

Dieser war lediglich als normales E-Fallseminar im Stundenplan vermerkt und nicht als besonderer Ausgangstest angekündigt, um eine gezielte Vorbereitung auf die Prüfungsinhalte zu verhindern. Der Ausgangstest beinhaltete dieselben 30 *Key-Feature*-Items wie der Eingangstest und sollte den kurzfristigen Lernerfolg der Studierenden unmittelbar nach der Lernphase messen.

Im darauffolgenden 4. klinischen Semester fand ein unangekündigter formativer Retentionstest statt, um den mittelfristigen Lernerfolg zu prüfen, ohne dass die Studierenden sich erneut gezielt auf die Prüfungsinhalte vorbereiteten. Dieser Retentionstest beinhaltete, analog zu ähnliche Studien (Larsen et al. 2009; Dobson und Linderholm 2015; Raupach et al. 2016), dieselben Items wie der Eingangs- und Ausgangstest und fand im Rahmen eines E-Fallseminars des Moduls 4.3 („Erkrankungen der Verdauungsorgane, des endokrinen Systems und des Stoffwechsels“) ca. fünf Monate nach dem jeweiligen Ausgangstest statt.

Die 30 *Key-Feature*-Items von Eingangs-, Ausgangs- und Retentionstest sind in Tabelle 1 zusammengefasst. Davon waren die Hälfte der Items Interventionsitems der Gruppe A und damit Kontrollitems der Gruppe B. Dies bedeutet, dass die Studierenden der Gruppe A diese 15 Items während der Lernphase ausschließlich im Videoformat bearbeiteten, im Gegensatz zu den Studierenden der Gruppe B, welche dieselben Items ausschließlich im Textformat bearbeitet haben. Bei der zweiten Hälfte der Items verhielt es sich genau entgegengesetzt. Diese Items waren Kontrollitems der Gruppe A und Interventionsitems der Gruppe B. Je nach Gruppenzugehörigkeit der Studierenden war ein einzelnes Item im Eingangs-, Ausgangs- oder Retentionstest also Interventions- oder Kontrollitem. Das *Key-Feature*-Item „Blutgasanalyse bei Verdacht auf beginnende CO_2-Narkose“ (*Key-Feature*-Fall zwei, *Key-Feature*-Item fünf) war beispielsweise ein Interventionsitem der Gruppe A und wurde dieser Gruppe während der Lernphase ausschließlich im Videoformat präsentiert, während es für Gruppe B demnach ein Kontrollitem war und dieser Gruppe ausschließlich im Textformat präsentiert wurde.

Tabelle 1: *Key-Feature*-Items von Eingangs-, Ausgangs- und Retentionstest und Gruppenzuordnung bezüglich Interventions- und Kontrollitems

Key Feature		***Key Feature***	**Interventionsitem**	**Kontrollitem**
Fall	**Item**			
1	**1**	Ableiten der Verdachtsdiagnose „Akute Lungenembolie“ bei typischer Klinik und Risikoprofil	B	A
	2	Wells-Score zur Entscheidung für oder gegen eine D-Dimer-Bestimmung	B	A
	3	Sofortiges Thorax-CT bei hoher Wahrscheinlichkeit und klinischer Stabilität	B	A
	4	Echokardiographie oder Troponinbestimmung zur Risikostratifizierung	B	A
	5	Indikation zur Fibrinolyse bei hämodynamischer Instabilität	B	A
	6	Erkennen eines nephrotischen Syndroms anhand der Labor- und Urinbefunde	A	B
	7	V. a. SLE bei typischer Klinik und wegweisendem AK-Befund	A	B
	8	Nierenbiopsie zur Diagnosesicherung	A	B
	9	Behandlung der Lupus-Nephritis durch einen Steroidstoß	A	B
2	**1**	Verdachtsdiagnose COPD bei chronischem Husten und Auswurf bei einem starken Raucher mit exspiratorischem Giemen	A	B
	2	Erkennen einer Obstruktion in der Lungenfunktion	A	B
	3	Erkennen einer links-apikalen Lobärpneumonie im Röntgenbild	A	B
	4	CRB-65-Index zur Entscheidung über die Empfehlung zur stationären Aufnahme	A	B
	5	Blutgasanalyse bei V. a. beginnende CO_2-Narkose	A	B
	6	NIV bei nachgewiesener Hyperkapnie	A	B
3	**1**	Verdacht auf sekundäre Hypertonie bei Manifestation >60 J., diskrepanten Organschäden und Therapierefraktärität	B	A
	2	Verdacht auf diastolische Dysfunktion bei guter EF, NYHA II und erhöhtem BNP	B	A
	3	ACE-Hemmer bei V. a. ACE-Hemmer-Husten absetzen	B	A
	4	Krankenhaus-Einweisung bei Hyponatriämie <120 mmol/l	B	A
	5	Thiaziddiuretika als mögliche Auslöser einer Hyponatriämie	B	A
	6	Zentrale pontine Myelinolyse als Komplikation einer zu raschen Natrium-Substitution	B	A
4	**1**	Schellong-Test bei V. a. orthostatisch bedingte Synkope	B	A
	2	Schädel-CT zur Diagnose eines Hirninfarktes	B	A
	3	Erkennen einer Tachyarrhythmia absoluta im EKG	B	A
	4	CHA_2DS_2-VASc-Score zur Indikationsstellung zur dauerhaften Antikoagulation	B	A
	5	Laborchemische Diagnose einer manifesten Hyperthyreose nach Konstrastmittelgabe	A	B
	6	Absetzen von Amiodaron bei manifester Hyperthyreose	A	B
	7	Klinischer Verdacht auf Lungenfibrose bei Dyspnoe und inspiratorischem „*Velcro*“-Knistern	A	B
	8	Amiodaron als Verursacher einer Lungenfibrose	A	B
	9	Indikationsstellung zur LTOT aufgrund eines pO_2 <55 mmHg in der arteriellen BGA	A	B

Legende: *Key-Feature*-Fälle und -Items von Raupach, erstmalige Verwendung an der UMG im Wintersemester 2007/08 (Raupach et al. 2009), fortlaufende Aktualisierung und Ergänzung, letztmalige Verwendung der *Key-Feature*-Fälle und -Items vor der Verwendung in der vorliegenden Arbeit in Raupach et al. (2016) und Andresen (2020).

Key-Feature-Items, welche im Eingangs-, Ausgangs- und Retentionstest geprüft wurden, wurden in der Lernphase zweimal geprüft. Bei erstmaligem Auftreten eines *Key-Feature*-Items sowie dessen Wiederholung handelte es sich zwar um dieselben Schlüsselstellen klinischer Entscheidungsfindung, jedoch in unterschiedlichen Fallbeispielen und Kontexten, sowohl im Vergleich zu Eingangs-, Ausgangs- und Retentionstest als auch zwischen den Erscheinungen. Tabelle 2 fasst den Verlauf der Wiederholungen zusammen.

Die im Rahmen dieser Arbeit durchgeführten Studien wurden von der Ethik-Kommission der UMG für das Wintersemester 2014/15 unter der Antragsnummer 11/9/14 und für das Sommersemester 2015 unter der Antragsnummer 22/4/15 genehmigt.

2.1.1 Implementierungs- und Pilotierungsphase Wintersemester 2014/15

Die erste Durchführung dieser Studie im Wintersemester 2014/15 diente als Implementierungs- und Pilotierungsphase. Neben der fortlaufenden Erstellung des Videomaterials für spätere E-Fallseminare diente der erste Durchlauf der Sammlung von Erfahrung im Umgang mit dem neuen Präsentationsformat und dem Aufdecken von möglichen Problemen seitens der technischen Implementierung. Im Eingangs-, Ausgangs- und Retentionstest wurden für beide Gruppen identisch nur Text-*Key-Feature*-Items verwendet. Erste Analysen der Daten ermöglichten eine Anpassung des Studiendesigns, die zur Beantwortung neu aufgeworfener Forschungsfragen in der zweiten Durchführung der Studie im Sommersemester 2015 führen sollten. Die angestellten statistischen Analysen hatten explorativen Charakter. Eingangs-, Ausgangs- sowie Retentionstest fanden am 22.10.2014, 23.01.2015 und 26.05.2015 statt.

Tabelle 2: Wiederholungen der *Key-Feature*-Items im Verlauf der Semesterwochen

Key Feature		**Semesterwoche**												
Fall	**Item**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	
1	1	Eingangstest						X		X			Ausgangstest	Retentionstest
	2							X		X				
	3							X		X				
	4							X		X				
	5							X		X				
	6								X		X			
	7								X		X			
	8								X		X			
	9								X		X			
2	1						X		X					
	2						X		X					
	3						X				X			
	4						X				X			
	5						X		X					
	6						X		X					
3	1			X		X								
	2			X		X								
	3			X						X				
	4									X		X		
	5									X		X		
	6									X		X		
4	1					X		X						
	2							X				X		
	3					X				X				
	4									X		X		
	5				X		X							
	6				X		X							
	7								X		X			
	8								X		X			
	9								X		X			

Legende: *Key-Feature*-Fall und -Item beziehen sich jeweils auf Tabelle 1

2.1.2 Studienphase Sommersemester 2015

Die zweite Durchführung der Studie im Sommersemester 2015 diente der Beantwortung der Forschungsfragen. Die hier gewonnenen Daten wurden im Rahmen dieser Arbeit einer ausführlichen Analyse unterzogen.

Während in der Implementierungs- und Pilotierungsphase im Wintersemester 2014/15 der Eingangs-, Ausgangs- und Retentionstest in identischer Form im Textformat verwendet wurde, wurden für die zweite Durchführung der Studie der Ausgangs- sowie Retentionstest verändert. Dies war nötig, um die Forschungsfrage Nr. 2, welche Rolle die Kongruenz zwischen Interventions- und Prüfungsformat spielt, beantworten zu können. Aus den bisher 30 Items dieser Tests wurden je ein Interventionsitem aus der Gruppe A und eines aus der Gruppe B eliminiert, um die Items in vier gleichgroße Gruppen einteilen zu können. Dabei wurden zwei eher leichte Items, die in der Implementierungs- und Pilotierungsphase hohe Itemschwierigkeiten (0,86 und 0,80) aufwiesen, ausgewählt. Hierbei handelte es sich um *Key-Feature*-Item neun aus dem ersten *Key-Feature*-Fall („Behandlung der Lupus-Nephritis durch einen Steroidstoß“) sowie *Key-Feature*-Item zwei aus dem vierten *Key-Feature*-Fall („Schädel-CT zur Diagnose eines Hirninfarktes“) (vgl. Tabelle 1). Von den verbliebenen 28 Items wurde nun sieben der 14 Interventionsitems von Gruppe A und sieben der 14 Interventionsitems von Gruppe B verfilmt. Durch diese weitere Ebene des Cross-overs konnten die 28 Items des Ausgangs- beziehungsweise Retentionstests nun in Abhängigkeit von der Gruppenzugehörigkeit eines Studierenden in die vier Kategorien „Lernphase: Text, Prüfungen: Text“ (1), „Lernphase: Text, Prüfungen: Video“ (2), „Lernphase: Video, Prüfungen: Text“ (3) und „Lernphase: Video, Prüfungen: Video“ (4) mit jeweils sieben Items eingeteilt werden. Eine Zuordnung der *Key-Feature*-Items zu entsprechenden Kategorien ist in der Tabelle 3 zusammengefasst. Eingangs-, Ausgangs- sowie Retentionstest fanden am 15.04.2015, 03.07.2015 und 11.12.2015 statt.

Tabelle 3: Gruppenzuordnung und Zuordnung Präsentationsformat der *Key-Feature*-Items im Ausgangs- und Retentionstest

Key Feature		**Interventionsitem**	**Kontrollitem**	**Präsentations-**	**Kategorie**	
Fall	**Item**	**Gruppe**	**Gruppe**	**format**	**Gruppe A**	**Gruppe B**
1	**1**	B	A	Text	1	3
	2	B	A	Video	2	4
	3	B	A	Text	1	3
	4	B	A	Video	2	4
	5	B	A	Video	2	4
	6	A	B	Video	4	2
	7	A	B	Text	3	1
	8	A	B	Text	3	1
2	**1**	A	B	Text	3	1
	2	A	B	Video	4	2
	3	A	B	Text	3	1
	4	A	B	Video	4	2
	5	A	B	Video	4	2
	6	A	B	Text	3	1
3	**1**	B	A	Text	1	3
	2	B	A	Video	2	4
	3	B	A	Video	2	4
	4	B	A	Video	2	4
	5	B	A	Text	1	3
	6	B	A	Text	1	3
4	**1**	B	A	Text	1	3
	3	B	A	Video	2	4
	4	B	A	Text	1	3
	5	A	B	Video	4	2
	6	A	B	Text	3	1
	7	A	B	Video	4	2
	8	A	B	Text	3	1
	9	A	B	Video	4	2

Legende: *Key-Feature*-Fall und -Item beziehen sich jeweils auf Tabelle 1. Kategorien: „Lernphase: Text, Prüfungen: Text“ (1), „Lernphase: Text, Prüfungen: Video“ (2), „Lernphase: Video, Prüfungen: Text“ (3) und „Lernphase: Video, Prüfungen: Video“ (4)

2.1.3 Probandenrekrutierung

Als mögliche Probanden für diese Studie wurden die Studierenden der Humanmedizin an der UMG im jeweils 3. klinischen Semester herangezogen. Diese wurden vor Beginn der Lehrveranstaltungen des Semesters per E-Mail sowie in der Einführungsveranstaltung des Moduls 3.1 „Erkrankungen des Herz-Kreislauf-Systems und der Lunge“ vom Studienleiter über die Hintergründe und Ziele der Studie informiert. Am Termin des ersten E-Fallseminars wurden Probanden-Informationszettel und entsprechende Einverständniserklärungen an

die Studierenden verteilt und bei Teilnahme an der Studie ausgefüllt wieder eingesammelt. Die Teilnahme an der Studie war freiwillig und es entstand den Studierenden kein Nachteil, wenn sie nicht daran teilnahmen.

2.1.4 Curriculare Implementierung

Die E-Fallseminare waren mit einer Länge von je 45 Minuten angesetzt und fanden in den Räumen 431 und 434 des Digitalen Prüfungs- und Schulungszentrums (DiPS, siehe Kapitel 2.3.1) der UMG statt. Unabhängig von einer Teilnahme an der Studie waren die E-Fallseminare als Pflichtveranstaltungen im Curriculum und damit in den Stundenplänen der Studierenden in den Modulen 3.1 „Erkrankungen des Herz-Kreislauf-Systems und der Lunge“, 3.2 „Erkrankungen der Niere und des Urogenitalsystems“ und 3.3 „Erkrankungen des Blutes, des Knochenmarks und Grundlagen der Tumorerkrankungen“ verankert. Im gesamten Semester waren maximal zwei Fehltermine möglich. Die in den E-Fallseminaren präsentierten Fallbeispiele orientierten sich thematisch an den in der Vorwoche gelehrten Themen der jeweiligen Module und sind in Tabelle 4 dargestellt. Inhaltlich wurden demnach die Teilbereiche der Inneren Medizin Kardiologie, Pneumologie, Nephrologie, Hämatologie und Onkologie in den E-Fallseminaren behandelt.

Der Retentionstest fand im Folgesemester unangekündigt im Rahmen eines E-Fallseminars des Moduls 4.3 („Erkrankungen der Verdauungsorgane, des endokrinen Systems und des Stoffwechsels“) ca. fünf Monate nach dem jeweiligen Ausgangstest statt. Nach Ende des Moduls 3.3 fand bis zum Retentionstest keine weitere explizite curriculare Lehre bezüglich der in dieser Studie untersuchten Inhalte mehr statt.

Tabelle 4: Curriculare Implementierung

SW	Themen		Präsentationsformat	
	Curriculare Lehre	**E-Fallseminar**	**Gruppe A**	**Gruppe B**
1	Koronare Herzerkrankung (KHK)	Alle Themen	Eingangstest	
2	Herzinsuffizienz	KHK	Video	Text
3	Vitien	Herzinsuffizienz	Text	Video
4	Arrhythmien	Vitien, KHK	Video	Text
5	Lungenerkrankungen	Arrhythmien, Herzinsuffizienz	Text	Video
6	Gefäßerkrankungen, Lungenembolie, entzündliche Herzerkrankungen	Lungenerkrankungen, Vitien, KHK	Video	Text
7	Nephrotisches Syndrom	Gefäßerkrankungen, Lungenembolie, entzündliche Herzerkrankungen, Arrhythmien	Text	Video
8	Homöostase	Nephrotisches Syndrom, Lungenerkrankungen	Video	Text
9	Nierenversagen	Homöostase, Gefäßerkrankungen, entzündliche Herzerkrankungen, Arrhythmien, Lungenembolie	Text	Video
10	Bluterkrankungen, Anämie	Nierenversagen, Nephrotisches Syndrom, Lungenerkrankungen	Video	Text
11	Lymphatische Erkrankungen	Arrhythmien, Homöostase, Bluterkrankungen, Anämie, Gefäßerkrankungen, Nierenversagen	Text	Video
12	Solide Tumore	Alle Themen	Ausgangstest	
FS	∅	Alle Themen	Retentionstest	

Legende: SW = Semesterwoche, FS = Folgesemester

2.2 Prüfungsmaterialien

Für diese Studie wurden *Key-Feature*-Fälle sowohl im Textformat als auch im Videoformat benötigt. Im Folgenden sollen die Entstehung und Produktion dieser Prüfungsmaterialien beschrieben werden.

2.2.1 *Key-Feature*-Fälle

Die in dieser Studie verwendeten *Key-Feature*-Fälle wurden von den Modulkoordinatoren sowie den lehrenden Ärzten der Module 3.1, 3.2 und 3.3 konstruiert und geschrieben. Überwiegend wurden diese bereits in vorherigen Semestern ausführlich an der UMG pilotiert und eingesetzt, neu erstellte *Key-Feature*-Fälle wurden in der Implementierungsphase pilotiert. Diese Fälle, die auf der Basis

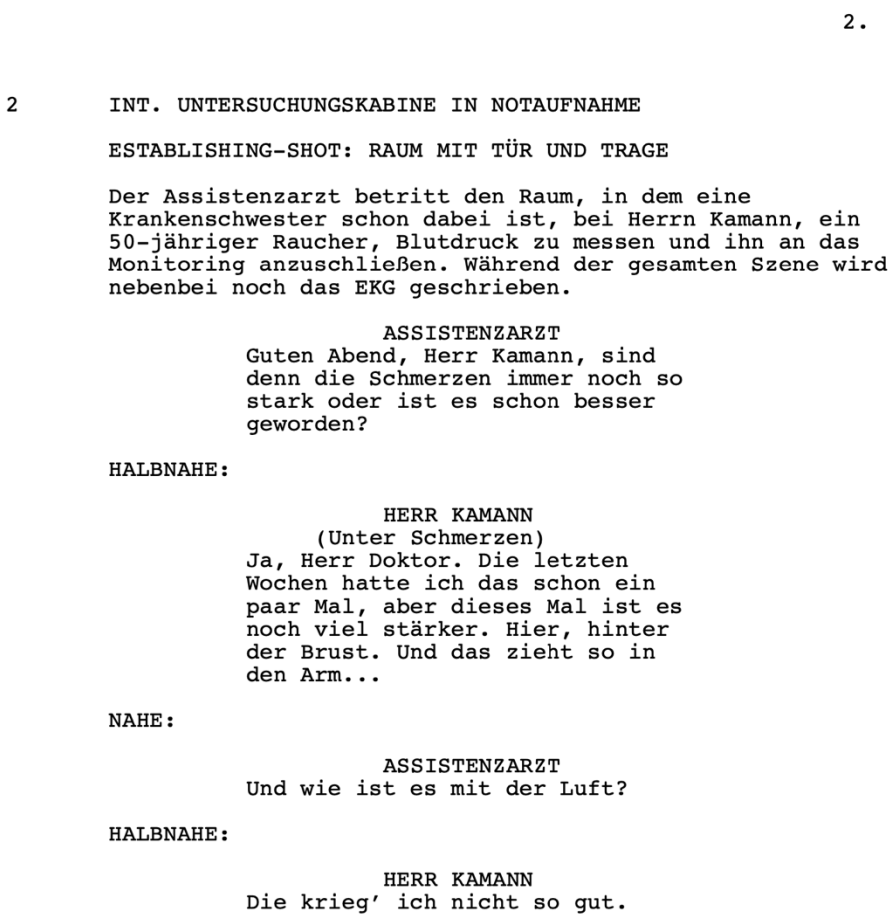
der in Vorstudien erhobenen Daten überarbeitet und aktualisiert worden waren, dienten als Grundlage der *Key-Feature*-Fälle für diese Studie.

Zur Beantwortung der *Key-Feature*-Fragen im *Long-Menu*-Format sind Antwortlisten mit sowohl den richtigen Antworten als auch mit Distraktoren notwendig. In dieser Studie wurden vier verschiedene Antwortlisten der Kategorien „Diagnosen“ (5621 Einträge), „Labor“ (757 Einträge), „technische Untersuchungen“ (751 Einträge) und „Therapieformen“ (1677 Einträge) genutzt. Diese Listen sind das Resultat aus Pilotierung und bisherigem Einsatz der *Key-Feature*-Fälle an der UMG und enthalten jeweils die in den Klammern genannte Anzahl von Einträgen. Anhand der Auswertungen der von den Studierenden gegebenen Antworten sowie Rückmeldungen der Studierenden wurden diese Listen stetig aktualisiert und ergänzt. Enthalten sind jeweils neben dem eigentlichen Begriff auch verschiedene Schreibweisen, Synonyme und gegebenenfalls Lokalisationen (z. B. „kranial“, „apikal“, „anterior“), um ein breites Spektrum an potentiellen Antwortmöglichkeiten abzudecken. So enthält beispielsweise die Antwortliste „Diagnosen“ neben dem Begriff „Myokardinfarkt“ auch die Begriffe „Myokardinfarkt, akut“ und „akuter Myokardinfarkt“ aber auch „Herzinfarkt“, „Herzinfarkt, akut“ und „akuter Herzinfarkt“ sowie „Herzinfarkt, anterior“, „Herzinfarkt apikal“, „STEMI“, „NSTEMI“ und viele weitere Begriffe. Die technische Umsetzung des *Long Menus* in dieser Studie zeigen die Bildschirmfotos in Abbildungen A10 f.

Nach Beantwortung jedes *Key-Feature*-Items hatten die Studierenden die Möglichkeit sich ein Feedback zu der entsprechenden Frage anzeigen zu lassen. Dies bestand zum einen aus einer kurzen Begründung der richtigen Antwort oder Erklärung häufiger Falschantworten sowie einer Liste mit den als richtig gewerteten Antworten. Dieses Feedback konnte für einen Großteil der Fragen aus den vorangehenden Anwendungen der *Key-Feature*-Fälle übernommen werden. Bereits bestehendes Feedback wurde vom Studienleiter aktualisiert und überarbeitet, noch nicht vorhandenes Feedback wurde vom Studienleiter neu verfasst. Das Feedback war nach jedem *Key-Feature*-Item als Bild eingefügt und wurde den Studierenden als Vorschau angezeigt und vergrößerte sich, wenn gewünscht, bei einem Klick auf die Vorschau. Die Einbindung des Feedbacks zeigen die Bildschirmfotos in Abbildung A2 ff. im Anhang.

2.2.2 Erstellung des Videomaterials

Die zu verfilmenden *Key-Feature*-Fälle mussten zunächst in Drehbücher umgeschrieben werden, um so die Fälle für das Medium Film vorzubereiten. Dabei mussten die in den Text-*Key-Feature*-Fällen gegebenen Situationsbeschreibungen audiovisuell dargestellt werden und die gegebenen medizinischen Informationen über die Patienten im Film in Dialogen zwischen Arzt und Patient, Studierendem, Pflegekraft oder weiteren Ärzten erarbeitet werden, um für den Zuschauer erlebbar zu sein. Insgesamt waren vier Personen an der Erstellung der Drehbücher mit der Software *Celtx* anhand der vorliegenden Text-*Key-Feature*-Fälle beteiligt. Abbildung 2 zeigt ein Beispiel einer Drehbuchseite.



2.

2 INT. UNTERSUCHUNGSKABINE IN NOTAUFNAHME

ESTABLISHING-SHOT: RAUM MIT TÜR UND TRAGE

Der Assistenzarzt betritt den Raum, in dem eine Krankenschwester schon dabei ist, bei Herrn Kamann, ein 50-jähriger Raucher, Blutdruck zu messen und ihn an das Monitoring anzuschließen. Während der gesamten Szene wird nebenbei noch das EKG geschrieben.

ASSISTENZARZT
Guten Abend, Herr Kamann, sind denn die Schmerzen immer noch so stark oder ist es schon besser geworden?

HALBNAHE:

HERR KAMANN
(Unter Schmerzen)
Ja, Herr Doktor. Die letzten Wochen hatte ich das schon ein paar Mal, aber dieses Mal ist es noch viel stärker. Hier, hinter der Brust. Und das zieht so in den Arm...

NAHE:

ASSISTENZARZT
Und wie ist es mit der Luft?

HALBNAHE:

HERR KAMANN
Die krieg' ich nicht so gut.

Abbildung 2: Bildschirmfoto einer Drehbuchseite

Zur Qualitätssicherung wurden die geschriebenen Drehbücher vor der Freigabe zum Dreh vom Studienleiter, einem weiteren Arzt und einer wissenschaftlichen Mitarbeiterin Korrektur gelesen, mit den Text-*Key-Feature*-Fällen abgeglichen und gegebenenfalls Korrekturen und Anpassungen vorgenommen.

Die Videoaufnahmen wurden mit verschiedenen Kameras der Firmen *Panasonic* und *Canon* auf Stativen gemacht. Für die Tonaufnahmen wurden ein Audiorekorder der Firma *Denon* sowie ein Audiorekorder der Firma *Zoom* verwendet. Von der Firma *Sennheiser* kamen das Mikrofon und die Mikrofonspeisung an einer Tonangel mit einem Mikrofon-Windschutz zum Einsatz. Die Ausleuchtung der Szenen wurde mithilfe von zwei Scheinwerfern der Firma *Ianiro* vorgenommen. Neben der Nutzung von im Studiendekanat vorhandenen Geräten, konnten weitere Teile der Filmausrüstung vom Videoteam der Niedersächsischen Staats- und Universitätsbibliothek Göttingen ausgeliehen werden. Ein Teil der Ausstattung wurde für das Projekt neu angeschafft. Eine detaillierte Auflistung der verwendeten Geräte befindet sich in Tabelle A2 im Anhang.

Die Schauspieler für die Filmaufnahmen wurden durch persönliche Ansprache gewonnen. Bei der Verteilung der Rollen wurde darauf geachtet, dass ärztliche Rollen, soweit möglich, auch von Ärzten der entsprechenden Fachrichtungen übernommen wurden. Weiterhin wirkten zahlreiche Personen mit medizinischer Expertise (Gesundheits- und Krankenpfleger, Medizinische Fachangestellte, Studierende der Medizin höherer Semester) in den Filmen mit. Als Schauspielpatienten konnten zum Großteil Laienschauspieler der Göttinger Lehrer-Theatergruppe „Lehrgut“ sowie aus dem „Theater im OP“ gewonnen werden. Es wurde darauf geachtet, dass das Alter der Schauspielpatienten in etwa mit dem Alter der fiktiven Patienten übereinstimmt, um eine realitätsnahe Darstellung der Fallbeispiele zu erreichen. Insgesamt wirkten 96 Personen als Schauspieler mit, davon unter anderem 28 Ärzte, 25 Schauspielpatienten und 29 Studierende. Externe Mitwirkende erhielten eine Aufwandsentschädigung.

Eine große Zahl von Szenen wurde auf dem Gelände der UMG gedreht. Dafür konnten einzelne kurzzeitig unbelegte Patientenzimmer und Flure auf den Stationen, Konferenzräume und Büros, Räume der kardiologischen, onkologischen und nephrologischen Ambulanzen und des ambulanten Palliativdienstes sowie das STÄPS, der Triage-Raum der Interdisziplinären Notaufnahme und das Lehr- und Simulationszentrum der Klinik für Anästhesiologie genutzt werden. Aber auch Funktionsbereiche wie die Echokardiografie oder das Herzkatheterlabor dienten für eine möglichst große Realitätsnähe als Kulisse. Alle Dreharbeiten auf dem Gelände der UMG wurden mit der Stabsstelle Unternehmens-

kommunikation, Presse- und Öffentlichkeitsarbeit abgestimmt. Neben der UMG wurde in vier verschiedenen Arztpraxen, vier Privatwohnungen, zwei Rettungs- beziehungsweise Krankentransportwagen sowie im Freien auf verschiedenen Wegen und Parkplätzen gedreht. Zur Rücksichtnahme auf die Privatsphäre der Patienten und die allgemeine Patientenversorgung wurden die Dreharbeiten in der Regel nach Beendigung der täglichen klinischen Arbeiten durchgeführt. Die Aufnahmen im Herzkatheterlabor entstanden beispielsweise während einer Wartung des Katheterlabors.

Zur Postproduktion wurden die aufgenommenen Video- und Audiodateien zunächst von den jeweiligen SD-Speicherkarten („Secure Digital"-Speicherkarte) auf den Computer in die Videoschnitt-Software *Final Cut Pro X* importiert und in sogenannten Mediatheken und Ereignissen nach den *Key-Feature*-Fällen sortiert und organisiert. Zunächst wurden die jeweils simultan aufgenommenen Video- und Audiodateien zu sogenannten Multicam-Clips zusammengeführt und automatisch, und falls dies keine zufriedenstellenden Ergebnisse lieferte manuell, anhand der Audiospuren synchronisiert. Dazu wurden in den Audiowellenformen (die grafische Darstellung der Audiospuren), die charakteristisch hohen Ausschläge des Klappe-Schlagens bzw. des Klatschens zu Beginn jeder Aufnahme aufgesucht und übereinandergelegt.

Im nächsten Schritt galt es bei einer ersten Sichtung der Clips einen Überblick über das zur Verfügung stehende Material zu gewinnen und verwertbare Aufnahmen und Einstellungen zu markieren. „Verwertbar" bezieht sich auf verschiedene Qualitätsmerkmale eines Videos wie die Übereinstimmung und Vollständigkeit des dargestellten Inhalts zu dem Drehbuch, der schauspielerischen Leistung der Darsteller, der akustischen Wahrnehmbarkeit des aufgenommenen Tons und einer ansprechenden visuellen Gestaltung der Einstellungen. Durch Aneinanderreihung der gewünschten markierten Aufnahmen konnte dann eine erste Arbeitsversion im Arbeitsbereich von *Final Cut Pro*, der sogenannten *Timeline* („Zeitleiste"), erstellt werden.

Die weitere Videobearbeitung beinhaltete das Kürzen oder Verlängern der gewählten Aufnahmen und das Setzen von Schnitten zwischen den verschiedenen Einstellungen eines Multicam-Clips. Waren einzelne Teile einer sonst in einer Aufnahme am besten gelungenen Szene in einer anderen Aufnahme der

gleichen Szene qualitativ höherwertig, so wurden diese punktuell ersetzt. Um einen nahtlosen und für den Zuschauer möglichst unmerkbaren Übergang zu erhalten wurde mit Überlappung, Ein- und Ausblenden der Audiospuren der verschiedenen Aufnahmen sowie entsprechender Schnittführung gearbeitet. Wo notwendig wurden Blenden, Texteinblendungen, Bilder von Elektrokardiogrammen, Röntgenaufnahmen oder Laborbefunden hinzugefügt. Erstere dienten der Visualisierung und Konkretisierung von Zeitsprüngen („Eine Woche später“), Letztere lieferten für den Fall relevante medizinische Befunde und Informationen, welche teilweise auch zur Beantwortung der Fragen des jeweiligen *Key-Feature*-Items notwendig waren. Am Ende jedes Video-*Key-Feature*-Items wurde die Frage des Items als Text eingeblendet. Bei Bedarf wurde eine Farbkorrektur der Videos vorgenommen um das Aussehen der verschiedenen Einstellungen, die mit verschiedenen Kameras gefilmt wurden, untereinander anzugleichen. Selten wurden weitere visuelle Effekte hinzugefügt, um die Handlung oder die Situation zu unterstreichen.

Die Audiospuren wurden mit Ein- und Ausblenden versehen und der Lautstärkepegel von verschiedenen Spuren angeglichen. Bei gleichzeitiger Begrenzung des Lautstärkepegels zur Verhinderung eines Übersteuerns des Audiosignals, konnte durch Einsatz eines Begrenzer-Effekts die gesamte Lautstärke eines Clips mit der *Gain*-Funktion angehoben werden. Dies war für eine bessere akustische Wahrnehmung des oft zu leisen Tons in der Regel notwendig. Wie bei den Videospuren wurde auch bei den Audiospuren gelegentlich mit Effekten und Umgebungsgeräuschen gearbeitet, um Situationen zu verdeutlichen. Unter anderem wurde bei Telefonaten eine Verzerrung der Stimme oder bei verbalisierten Gedankengängen einer Person ein Nachhall-Effekt eingesetzt.

So entstanden durch Videobearbeitung die 164 geplanten Video-*Key-Feature*-Items mit einer Gesamtspieldauer von 03 h 05 min beziehungsweise einer mittleren Länge von 68 Sekunden je Item. Diese wurden aus *Final Cut Pro* in einem MP4-Dateicontainer (Auflösung: 720p; Videocodec: H.264; Audiocodec: MP3) exportiert. Mit der frei verfügbaren Software *HandBrake* wurden die exportierten Videos in demselben Format und derselben Auflösung mit einer geringeren Datenrate recodiert um die Dateigröße zu reduzieren und Kompatibilität mit dem Prüfungssystem herzustellen.

Zur Qualitätssicherung wurden die Videos vor der Freigabe zur Verwendung von dem Studienleiter, einem weiteren Arzt und einer wissenschaftlichen Mitarbeiterin angesehen und eventuelle Korrekturen umgesetzt. Informationen zur Verfügbarkeit des Prüfungsmaterials befinden sich im Anhang.

2.3 Ablauf der E-Fallseminare

Im Folgenden wird die Einbindung der fertigen Videos in das Prüfungssystem zur Verwendung in den E-Fallseminaren sowie deren Durchführung beschrieben.

2.3.1 Das Digitale Prüfungs- und Schulungszentrum der UMG

Das Digitale Prüfungs- und Schulungszentrum der UMG wurde zum Wintersemester 2012/13 in Betrieb genommen und besteht aus den drei Räumen 431, 434 und 257, von denen im Rahmen dieser Studie nur die ersten beiden genutzt wurden. Jeder dieser beiden Räume hat 77 Arbeitsplätze mit sogenannten *Thin-Computer-Clients* mit Bildschirm, Tastatur, Maus und Kopfhörern. Diese *Clients* verfügen nur über eine minimale Hardware und dienen als Ein- und Ausgabegeräte für leistungsstärkere virtuelle Computer, welche sich auf Servern im Rechenzentrum des Geschäftsbereichs Informationstechnologie der UMG befinden. Eine Software, die sogenannte Klassenraumsteuerungssoftware *NetMan*, ermöglicht die zentrale Steuerung der 154 *Clients*. So können bestimmte Funktionen wie zum Beispiel die Internetnutzung und Software während der E-Fallseminare auf den *Clients* eingeschränkt oder deaktiviert werden.

2.3.2 Prüfungserstellung für die E-Fallseminare

Die für die Durchführung der E-Fallseminare notwendigen Prüfungen wurden mit der webbasierten Software *Item Management System* erstellt. Zunächst wurde eine neue Prüfung angelegt und konfiguriert. Benötigte Medien wie Bilder und Videos wurden im *Item Management System* angelegt und hochgeladen. Die einzelnen Items der Prüfung wurden erstellt und die Texte, Medien und Fragen eingefügt. Für die *Key-Feature*-Fragen wurden die entsprechenden *Long-Menu*-Antwortlisten mit den verschiedenen richtigen Antworten im System definiert. Für jedes E-Fallseminar der Lernphase wurden zwei verschiedene

Versionen der Prüfung erstellt, eine im Textformat und eine im Videoformat. Die fertiggestellten Prüfungen wurden anschließend im QTI-Datenformat (*Question & Test Interoperability*, ein standardisiertes Datenformat für E-Prüfungen) aus dem *Item Management System* exportiert und heruntergeladen.

2.3.3 Durchführung eines E-Fallseminars

Die QTI-Dateien mit den Prüfungen im Text- sowie Videoformat für ein E-Fallseminar wurden dann im Computersystem des DiPS über die *Management-Console* des *CAMPUS-Prüfungssystems* in dieses importiert. In einem nächsten Schritt mussten für jede Prüfung mehrere PINs (Persönliche Identifikationsnummern) generiert werden, bevor die Prüfungen auf die 154 Computer-Clients, auf welchen der *CAMPUS-Prüfungsplayer* läuft, verteilt werden konnten. Während der Lernphase wurde dabei die Prüfung im Textformat auf die 77 Clients des einen Raumes und die Prüfung im Videoformat auf die 77 Clients des anderen Raumes des DiPS verteilt. Damit war die Vorbereitung eines E-Fallseminares abgeschlossen.

Vor dem Beginn jedes E-Fallseminars wurden die Studierenden beim Einlass in die Räume noch einmal von den Aufsichtführenden nach ihrer Gruppenzugehörigkeit gefragt, um trotz der eindeutigen Raumzuteilung in den Stundenplänen sicher zu stellen, dass sich jeder Studierende in dem ihm zugewiesenen Raum mit dem passenden Präsentationsformat für die Woche befand.

Die Studierenden konnten ihre Sitzplätze im zugeteilten Raum frei wählen. Um an der Prüfung teilzunehmen mussten die Studierenden sich mit ihrer studentischen Nutzerkennung und ihrem Passwort in einer Eingabemaske anmelden (vgl. Abbildung A1 im Anhang). Die Prüfung begann für alle Studierenden eines Raumes zeitgleich mit der Bekanntgabe der PIN durch die Aufsichtführenden, welche auf ein Whiteboard geschrieben wurde. Abbildung 3 zeigt die Durchführung eines E-Fallseminars im DiPS.

Abbildung 3: Foto der Durchführung eines E-Fallseminars im DiPS Wintersemester 2014/15

Die angezeigte Reihenfolge der *Key-Feature*-Fälle eines E-Fallseminars wurde vom *CAMPUS-Prüfungssystem* von Arbeitsplatz zu Arbeitsplatz zufällig permutiert, während die Reihenfolge der *Key-Feature*-Items innerhalb eines Falles aufgrund ihres logischen Aufbaus unverändert blieb. Die Reihenfolge des Bearbeitens der Fälle konnte von den Studierenden frei gewählt werden. Innerhalb eines *Key-Feature*-Falls mussten die Fragen jedoch der vorgegebenen Reihenfolge nach beantwortet werden. Erst nach Bestätigung der gegebenen Antwort ließ sich die nächste Frage aufrufen. Bereits beantwortete und bestätigte Fragen konnten zwar weiterhin angesehen, jedoch nicht mehr bearbeitet werden. Dies war notwendig um den Studierenden nach Beantwortung jedes *Key-Feature*-Items das in Kapitel 2.2.1 beschriebene Feedback zeigen zu können und gleichzeitig zu verhindern, dass die Antwort nach dem Studium des Feedbacks noch einmal geändert wurde. Die Bildschirmfotos in Abbildung A2 ff. im Anhang zeigen Beispiele eines Text-*Key-Feature*-Falls. Für die Video-*Key-Feature*-Items wurde eine Vorschau in Form eines Miniaturstandbildes aus dem Video im *CAMPUS-Prüfungsplayer* angezeigt. Nach einem Klick auf die Vorschau öffnete sich das Video in der Abspielsoftware *Microsoft Windows Media*

Player, wie in Abbildung 4 dargestellt. Weitere Bildschirmfotos eines Video-*Key-Feature*-Falls befinden sich in Abbildung A9 ff. im Anhang.

Abbildung 4: Bildschirmfoto der Darstellung eines Video-*Key-Feature*-Items im *Windows Media Player.* Durch Schließen des Fensters gelangten die Studierenden zurück in den *CAMPUS-Prüfungsplayer*.

Nach jedem der präsentierten *Key-Feature*-Fälle wurde den Studierenden eine thematisch zum jeweiligen *Key-Feature*-Fall oder zum aktuellen Modul passende alte Klausurfrage (klassische Einfachauswahlfrage) aus einer der vorangegangenen Modulklausuren des jeweiligen Moduls präsentiert. Dies sollte den Studierenden als weitere Vorbereitung auf die aktuelle Modulklausur einen Mehrwert bieten und als Anreiz dienen. Nach der Bearbeitung der *Key-Feature*-Fälle konnten die Studierenden das DiPS verlassen. Nachdem alle Studierenden mit der Bearbeitung fertig waren bzw. nach maximal 45 Minuten wurden die Prüfungen in beiden Räumen in der *Management-Console* des *CAMPUS-Prüfungssystems* beendet, die Prüfungsdaten im Excel-Dateiformat exportiert und an den Studienleiter übermittelt (vgl. Kapitel 2.5).

2.4 Evaluationsbogen

Um die studentische Sichtweise auf die E-Fallseminare und die beiden Präsentationsformate zu untersuchen, wurden die Probanden nach dem Ausgangstestat im Sommersemester 2015 gebeten, einen Evaluationsbogen auszufüllen.

Die Evaluation bestand aus einem Fragebogen, der sich aus den Teilen „Allgemeines“, „Inhalt“, „Feedback“, „Technische Umsetzung“ und „Abschluss“ zusammensetzte; sie wurde mit der Software *EvaSys* umgesetzt. Der anonyme Fragebogen enthielt neben Fragen zu Alter und Geschlecht überwiegend Items zu den E-Fallseminaren, Video-*Key-Features* und Feedback, welche die Studierenden auf einer sechsstufigen Likert-Skala von eins „trifft voll zu“ bis sechs „trifft überhaupt nicht zu“ bewerten sollten. Eine Einfachantwort-Frage und eine Gesamtbewertung nach dem Schulnotenprinzip ergänzten den Fragebogen.

2.5 Datensammlung und statistische Analyse

Zunächst wurden die nach jedem E-Fallseminar durch Mitarbeiter der Lehr-Informationstechnologie exportierten Excel-Dateien aller Erhebungszeitpunkte durch den Studienleiter auf mögliche richtige Antworten der Studierenden, die im System noch nicht als richtige Antworten definiert waren, untersucht und gegebenenfalls korrigiert. Zur statistischen Analyse wurden die einzelnen Datensätze zu einem longitudinalen Datensatz zusammengefügt, welcher dann anonymisiert wurde.

2.5.1 Dummy-Codierung

Vor der statistischen Analyse wurde eine sogenannte Dummy-Codierung durchgeführt. Dazu wurden jeweils für den Eingangs-, den Ausgangs- und den Retentionstest für jeden Probanden in neuen Variablen die Summe der Punkte aller Interventionsitems (Items, welche in der Lernphase im Videoformat bearbeitet wurden) sowie aller Kontrollitems (Items, welche in der Lernphase im Textformat bearbeitet wurden) gebildet. So war es möglich, in der statistischen Analyse die gesamte Stichprobe zu betrachten.

Zusätzlich wurden Punktsummen für weitere Variablen mit der Zuordnung der Items zu den vier Kategorien *„Lernphase: Text, Prüfungen: Text“*, *„Lernphase:*

Text, Prüfungen: Video“, *„Lernphase: Video, Prüfungen: Text“* und *„Lernphase: Video, Prüfungen: Video“* gebildet (vgl. Kapitel 2.1.2 sowie Tabelle 3). Daraus ließen sich weitere Summen für alle Items mit Kongruenz zwischen Präsentationsformat in der Lernphase und den formativen Prüfungen (*„Lernphase: Text, Prüfungen: Text“* und *„Lernphase: Video, Prüfungen: Video“*) sowie Items mit Inkongruenz zwischen diesen (*„Lernphase: Text, Prüfungen: Video“* und *„Lernphase: Video, Prüfungen: Text“*) bilden. Tabelle 5 zeigt eine Übersicht über die Dummy-Codierung der erreichten Punkte in den konsekutiven Prüfungen.

Weitere Dummy-Codierung wurde für andere vom Präsentationsformat abhängige Variablen, wie zum Beispiel der mittleren Bearbeitungszeit der E-Fallseminare, durchgeführt.

Tabelle 5: Übersicht der Kategorien der Dummy-Codierung

Nr.	Dummy-Codierung für	Punktsumme gebildet aus
1	„Lernphase*:* Text, Prüfungen*:* Text“	Items, die im Textformat gelernt und geprüft wurden
2	„Lernphase*:* Text, Prüfungen*:* Video“	Items, die im Textformat gelernt und im Videoformat geprüft wurden
3	„Lernphase*:* Video, Prüfungen*:* Text“	Items, die im Videoformat gelernt und im Textformat geprüft wurden
4	„Lernphase*:* Video, Prüfungen*:* Video“	Items, die im Videoformat gelernt und geprüft wurden
5	„Lernphase*:* Text“	Kontrollitems (Nrn. 1 und 2)
6	„Lernphase*:* Video“	Interventionsitems (Nrn. 3 und 4)
7	„Kongruente Formate“	Items, die im Textformat gelernt und geprüft wurden sowie Items, die im Videoformat gelernt und geprüft wurden (Nrn. 1 und 4)
8	„Inkongruente Formate“	Items, die im Videoformat gelernt und im Textformat geprüft wurden und Items, die im Textformat gelernt und im Videoformat geprüft wurden (Nrn. 2 und 3)

2.5.2 Quantitative Datenanalyse

Die statistischen Analysen in dieser Arbeit wurden mit dem Statistikprogramm *SPSS Statistics* durchgeführt. Das Signifikanzniveau wurde auf 5 % gesetzt. Sofern nicht anders beschrieben werden die Ergebnisse deskriptiver Analysen in Mittelwert ± Standardabweichung bzw. als Proportion in % (n) angegeben. Die Charakteristiken der Populationen wurden mithilfe deskriptiver Statistik ermittelt.

Der primäre Endpunkt der Studie war der Unterschied der Mittelwerte der von den Studierenden erreichten Punkte im Ausgangstest in Abhängigkeit von dem Präsentationsformat eines Items (Text oder Video) während der Lernphase als Maß für den kurzfristigen Lernerfolg. Analog dazu war der sekundäre Endpunkt der mittelfristige Lernerfolg der Studierenden zum Zeitpunkt des Retentionstests in Abhängigkeit von dem Präsentationsformat während der Lernphase. Eingeschlossen in die primäre und sekundäre Analyse wurden alle Studierenden, welche bei Eingangs-, Ausgangs- und Retentionstest anwesend waren sowie keine Kontamination im Studienverlauf erfahren, also nie in einer falschen Gruppe an E-Fallseminaren teilgenommen hatten.

Zunächst wurden ordinal und metrisch skalierte Daten mithilfe des Kolmogorow-Smirnow-Tests auf Normalverteilung überprüft. Um in der primären und sekundären Endpunktanalyse Unterschiede zwischen zwei Gruppen zu detektieren wurden – je nach Ergebnis der Testung auf Normalverteilung – gepaarte t-Tests oder Wilcoxon-Tests durchgeführt. Alle weiteren statistischen Analysen hatten explorativen Charakter.

Bei einer potentiellen Stichprobengröße von 125 Studierenden im Sommersemester 2015 konnte in dieser Studie ein Effekt der Stärke 0,25 (Cohen’s d (Cohen 1992)) mit einer Power von 80 % detektiert werden. Die Effektstärke von 0,25 wurde gewählt, da dies ein halb so großer Effekt wäre wie in der Vorstudie, in der repetitives Prüfen mit repetitivem Fallstudium verglichen wurde (Raupach et al. 2016). Der zusätzliche Effekt durch den Einsatz von Videos wurde vor dem Hintergrund der verfügbaren Literatur als geringer eingeschätzt.

Aufgrund einer kleinen erwarteten Effektstärke bei feststehender Stichprobengröße erfolgte zusätzlich eine Analyse nach Bayes, um eine Tendenz des erwarteten Effekts abschätzen zu können. Die Berechnung des Bayes-Faktors erfolgte mit dem Bayes-Faktor-Rechner nach Dienes (2008).

Zur Beantwortung der Forschungsfrage Nr. 2 nach Kongruenz zwischen Präsentationsformat in der Lernphase und den formativen Prüfungen wurde ein Friedman-Test durchgeführt, welchem sich bei Signifikanz ein Wilcoxon-Vorzeichen-Rang-Test anschloss, um Unterschiede zu detektieren.

Zur Beantwortung der 3. Forschungsfrage sollten wesentliche Charakteristika (Geschlecht, Alter, Prüfungsleistungen) als mögliche Prädiktoren eines besonders großen Lernerfolgs untersucht werden. Hierfür wurde für jeden Studierenden jeweils für das Text- und Videoformat die Differenz der erreichten Punkte des Ausgangs- und des Eingangstests als Wissenszuwachs berechnet. Anhand dieses Wissenszuwachses wurden die Studierenden in drei gleichgroße Gruppen (je $n = 31$) stratifiziert. In einem nachfolgenden Schritt wurden für jedes Präsentationsformat das Drittel der Studierenden mit dem höchsten Wissenszuwachs („oberes Drittel“) mit dem Drittel der Studierenden mit dem geringsten Wissenszuwachs („unteres Drittel“) bezüglich ihres Geschlechts (Chi^2-Test), ihres Alters und ihrer Klausurvorleistungen (t-Test für unabhängige Stichproben) verglichen.

Zur Beantwortung der Forschungsfrage Nr. 4 wurde zur Abschätzung der Reliabilität als Maßzahl für die interne Konsistenz der E-Fallseminare Cronbachs α berechnet. Die verwendeten Items wurden einer Itemanalyse unterzogen, in deren Rahmen Itemschwierigkeit und Trennschärfe berechnet wurden.

Die Ergebnisse der Evaluationsbögen zur Beantwortung der 5. Forschungsfrage wurden deskriptiv dargestellt.

3 Ergebnisse

3.1 Implementierungs- und Pilotierungsphase Wintersemester 2014/15

Die Ergebnisse der Durchführung der Studie im Wintersemester 2014/15 werden lediglich deskriptiv dargestellt, da es sich um die Implementierungs- und Pilotierungsphase vor der eigentlichen Studie handelte.

3.1.1 Deskriptive Analysen Wintersemester 2014/15

Von 112 in den Modulen 3.1, 3.2, 3.3 und 4.3 angemeldeten Studierenden gaben 106 ihr schriftliches Einverständnis, an der Studie teilzunehmen. Durch Kontamination während der Lernphase sowie Nichterscheinen zu Eingangs-, Ausgangs- oder Retentionstest mussten weitere 17 Studierende von der Auswertung ausgeschlossen werden.

Die gesamte Stichprobe bestand aus 89 Studierenden, davon waren 58,4 % (52) weiblich und 41,6 % (37) männlich. Das mittlere Alter am ersten Tag des Wintersemesters 2014/15 betrug 25,3 ± 3,2 Jahre, die im Vorsemester in den regulären Klausuren erreichte Leistung der Studierenden betrug im Mittel 85,3 ± 5,5 %.

Gruppe A bestand aus 46 Studierenden, davon waren 56,5 % (26) weiblich und 43,5 % (20) männlich. Das mittlere Alter betrug 24,6 ± 3,0 Jahre und die erreichte Leistung im Vorsemester 84,7 ± 5,8 %.

Gruppe B bestand aus 43 Studierenden, davon waren 60,5 % (26) weiblich und 39,5 % 17 männlich. Das mittlere Alter betrug 25,9 ± 3,3 Jahre und die erreichte Leistung im Vorsemester 85,9 ± 5,2 %.

Die Gruppen der eingeschlossenen Studierenden unterschieden sich weiterhin hinsichtlich Alter ($p = 0,060$), Geschlecht ($p = 0,706$) und bisheriger Studienleistung ($p = 0,359$) nicht signifikant voneinander.

3.1.2 Ergebnisse des Eingangs-, Ausgangs- und Retentionstestes Wintersemester 2014/15

Im Eingangstest erreichten die Studierenden im Mittel 30,4 ± 13,4 %, im Ausgangstest 71,6 ± 16,0 % und im Retentionstest 71,6 ± 16,0 %. Abbildung 5 zeigt diese Mittelwerte im zeitlichen Verlauf.

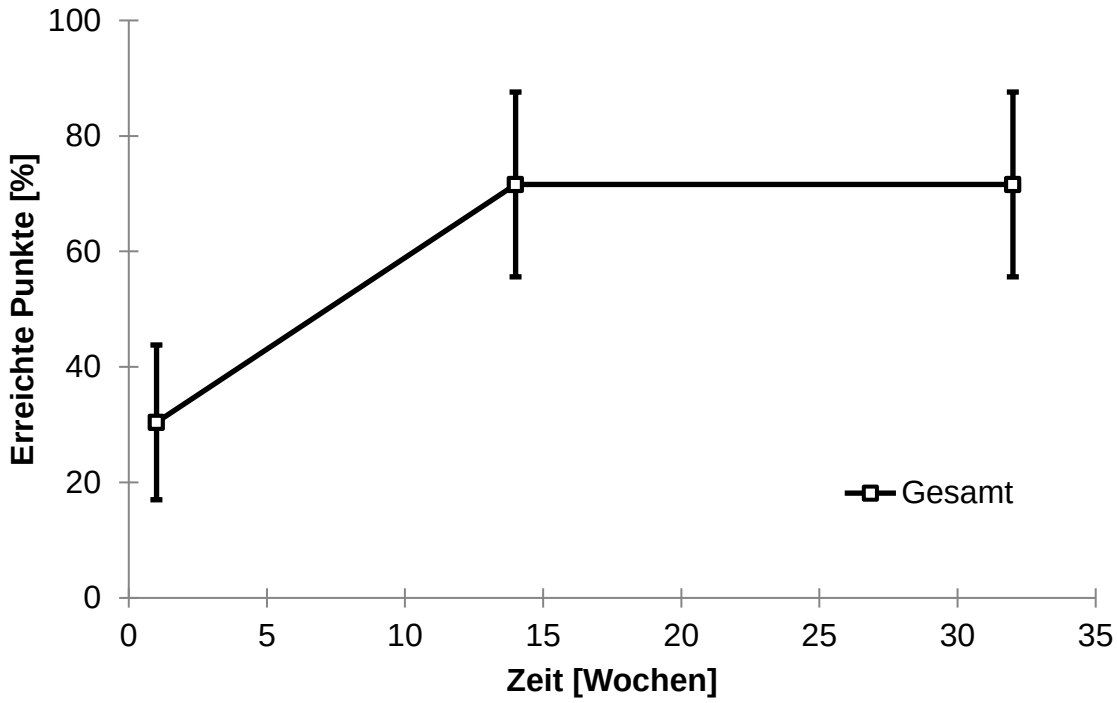


Abbildung 5: Durchschnittlich in Eingangs-, Ausgangs- und Retentionstest erreichte Punkte während der Pilotierung Wintersemester 2014/15 in %, unabhängig vom Präsentationsformat. Die Balken zeigen die Standardabweichung.

3.1.3 Itemanalyse und Testgütekriterien Wintersemester 2014/15

Im Rahmen der Pilotierung wurden für jedes E-Fallseminar und Präsentationsformat die Itemschwierigkeiten und Trennschärfen berechnet und gemittelt. Als Maßzahl für die interne Konsistenz wurde Cronbachs α je E-Fallseminar und Präsentationsformat berechnet. Für diese Berechnungen wurden jeweils die Daten aller Studienteilnehmer, die jeweils bei einem E-Fallseminar anwesend waren, herangezogen. Die Ergebnisse dieser Analysen sind in Tabelle 6 zusammengefasst.

Die insgesamt über die Lernphase gemittelten Itemschwierigkeiten der Video-*Key-Feature*-Items unterschieden sich nicht signifikant von denen der Text-*Key-*

Feature-Items (0,65 ± 0,07 vs. 0,64 ± 0,08). Dies gilt auch für die Trennschärfen (0,18 ± 0,08 vs. 0,17 ± 0,06) sowie für die Mittelwerte des Cronbachs α der E-Fallseminare im Videoformat gegenüber denen im Textformat (0,50 vs. 0,48). *Key-Feature*-Items mit unzulänglichen Itemkennwerten wurden für die Studienphase im Sommersemester 2015 teilweise überarbeitet und angepasst.

Tabelle 6: Cronbachs α und gemittelte Itemkennwerte der Pilotierung Wintersemester 2014/15

E-Fallseminar Semesterwoche		**Cronbachs α**	**Itemschwierigkeit**	**Trennschärfe**
Eingangstest		0,74	0,30 ± 0,25	0,25 ± 0,12
2	Text	0,48	0,68 ± 0,30	0,16 ± 0,15
	Video	0,49	0,65 ± 0,30	0,15 ± 0,16
3	Text	0,47	0,53 ± 0,31	0,16 ± 0,13
	Video	0,26	0,60 ± 0,31	0,08 ± 0,14
4	Text	0,39	0,74 ± 0,26	0,14 ± 0,18
	Video	0,35	0,76 ± 0,27	0,08 ± 0,14
5	Text	0,56	0,52 ± 0,28	0,22 ± 0,13
	Video	0,60	0,57 ± 0,26	0,23 ± 0,14
6	Text	0,24	0,77 ± 0,14	0,06 ± 0,15
	Video	0,34	0,79 ± 0,16	0,10 ± 0,18
7	Text	0,39	0,67 ± 0,17	0,12 ± 0,17
	Video	0,58	0,68 ± 0,18	0,21 ± 0,17
8	Text	0,66	0,66 ± 0,22	0,27 ± 0,16
	Video	0,64	0,63 ± 0,22	0,28 ± 0,14
9	Text	0,45	0,53 ± 0,18	0,16 ± 0,15
	Video	0,70	0,60 ± 0,20	0,30 ± 0,18
10	Text	0,67	0,65 ± 0,23	0,27 ± 0,14
	Video	0,42	0,69 ± 0,21	0,13 ± 0,18
11	Text	0,45	0,60 ± 0,23	0,16 ± 0,14
	Video	0,62	0,59 ± 0,23	0,25 ± 0,15
Ausgangstest		0,80	0,72 ± 0,17	0,30 ± 0,12
Retentionstest		0,80	0,72 ± 0,18	0,32 ± 0,12

3.2 Studienphase Sommersemester 2015

3.2.1 Deskriptive Analysen Sommersemester 2015

Von 125 in den Modulen 3.1, 3.2, 3.3 und 4.3 angemeldeten Studierenden gaben 95,2 % (119) ihr schriftliches Einverständnis, an der Studie teilzunehmen. Durch Kontamination während der Lernphase sowie Nichterscheinen zu Eingangs-, Ausgangs- oder Retentionstest mussten weitere 26 Studierende von der primären Auswertung ausgeschlossen werden, sodass sich eine Rückläuferquote von 74,4 % (93) ergibt. Die Abbildung 6 zeigt den Verlauf und die entsprechende Verteilung der ausgeschlossenen Teilnehmer. Es gab keinen Unterschied bezüglich Alter ($p = 0,197$), Geschlecht ($p = 0,816$) oder bisheriger Studienleistung ($p = 0,804$) zwischen ein- beziehungsweise ausgeschlossenen Studienteilnehmern.

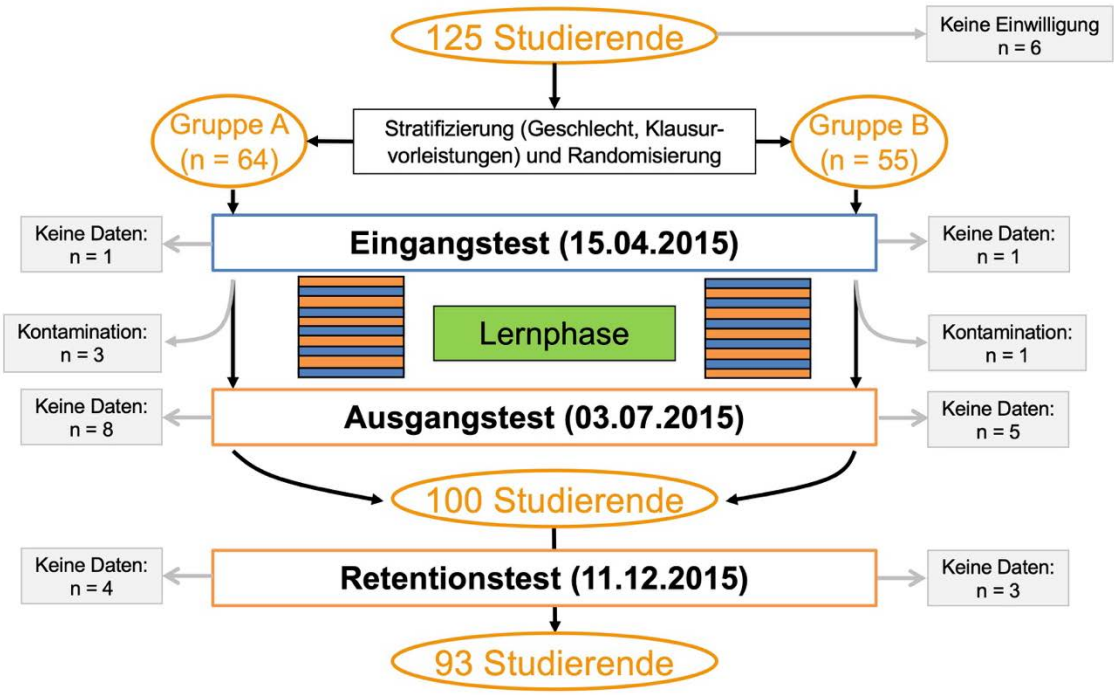


Abbildung 6: Schema des Studienverlaufs Sommersemester 2015 mit ein- und ausgeschlossenen Studierenden

Die gesamte Stichprobe für die Analyse des primären und sekundären Endpunkts bestand aus 93 Studierenden, davon waren 64,5 % (60) weiblich und 33,5 % (33) männlich. Das mittlere Alter am ersten Tag des Sommersemesters 2015 betrug 25,8 ± 3,9 Jahre, die im Vorsemester in den regulären Klausuren erreichte Leistung der Studierenden betrug im Mittel 82,2 ± 7,5 %.

Gruppe A bestand aus 58,3 % (28) weiblichen und 41,7 % (20) männlichen Studierenden mit einem mittleren Alter von 25,5 ± 2,8 Jahren und einer erreichten Klausurvorleistung von 81,4 ± 8,2 %.

Gruppe B bestand aus 71,1 % (32) weiblichen und 28,9 % (13) männlichen Studierenden mit einem mittleren Alter von 26,2 ± 4,8 Jahren und einer erreichten Klausurvorleistung von 82,9 ± 6,6 %.

Die Gruppen der eingeschlossenen Studierenden unterschieden sich hinsichtlich Alter ($p = 0,423$), Geschlecht ($p = 0,198$) und bisheriger Studienleistung ($p = 0,335$) nicht signifikant voneinander. Somit war die Randomisierung auch nach Ausschluss entsprechender Studienteilnehmer gegeben.

Die Anzahl der anwesenden Studierenden je E-Fallseminar kann der Tabelle 7 entnommen werden. Da die Teilnahme bei Eingangs-, Ausgangs- und Retentionstest Einschlusskriterium war, lag diese zu allen drei Zeitpunkten in beiden Gruppen bei 100 % und wird hier nicht gesondert aufgeführt.

Tabelle 7: Anwesenheit bei den E-Fallseminaren in % (n)

Semesterwoche	**Gesamt (n = 93)**	**Gruppe A (n = 48)**	**Gruppe B (n = 45)**
Woche 2	94,6 (88)	91,7 (44)	97,8 (44)
Woche 3	95,7 (89)	95,8 (46)	95,6 (43)
Woche 4	97,8 (91)	97,9 (47)	97,8 (44)
Woche 5	96,8 (90)	95,8 (46)	97,8 (44)
Woche 6	96,8 (90)	95,8 (46)	97,8 (44)
Woche 7	94,6 (88)	91,7 (44)	97,8 (44)
Woche 8	90,3 (84)	87,5 (42)	93,3 (42)
Woche 9	94,6 (88)	97,9 (47)	91,1 (41)
Woche 10	87,1 (81)	89,6 (43)	84,4 (38)
Woche 11	84,9 (79)	83,3 (40)	86,7 (39)
Mittelwert:	93,3 (86,8)	92,7 (44,5)	88,1 (42,3)

Insgesamt nahmen 96,8 % aller eingeschlossenen Studierenden an mindestens acht der zehn E-Fallseminare der Lernphase teil. Es wurden keine Unterschiede zwischen den beiden Studiengruppen gefunden.

Die mittlere Bearbeitungszeit eines E-Fallseminars der Lernphase mit seinen 15 *Key-Feature*-Items betrug 25:04 ± 04:24 min:s. Die Bearbeitungszeiten von

Eingangs-, Ausgangs- und Retentionstest waren aufgrund der größeren Anzahl von zu bearbeitenden Items (28 Items) entsprechend länger. Tabelle 8 zeigt die Mittelwerte der Bearbeitungszeiten der E-Fallseminare aufgeteilt nach Gruppen. Während der Lernphase brauchte diejenige Gruppe, welche in der jeweiligen Woche die *Key-Feature*-Fälle im Videoformat bearbeitete, jeweils signifikant länger als diejenige Gruppe, welche die Fälle im Textformat bearbeitete.

Zur Bearbeitung von E-Fallseminaren im Videoformat benötigten die Studierenden beider Gruppen im Mittel 31:15 ± 04:25 min:s, für E-Fallseminare im Textformat im Mittel 18:56 ± 04:32 min:s. Damit betrug die mittlere Zeit, die die Studierenden länger zur Bearbeitung der E-Fallseminare im Videoformat gegenüber dem Textformat benötigten, 12:18 ± 02:13 min:s.

Die Gruppe B benötigte mit im Mittel 32:13 ± 04:35 min:s länger zur Bearbeitung von E-Fallseminaren im Videoformat als Gruppe A mit 30:21 ± 04:06 min:s ($p = 0{,}042$). Weitere Gruppenunterschiede wurden nicht gefunden.

Tabelle 8: Bearbeitungszeit der E-Fallseminare in min:s

E-Fallseminar	Gruppe A (n = 48)	Gruppe B (n = 45)
Eingangstest	41:47 ± 07:50	39:25 ± 07:14
Semesterwoche 2	36:19 ± 05:25	24:26 ± 06:24
Semesterwoche 3	21:57 ± 05:49	38:44 ± 05:28
Semesterwoche 4	31:06 ± 03:37	19:16 ± 05:22
Semesterwoche 5	20:38 ± 06:19	34:49 ± 05:18
Semesterwoche 6	27:35 ± 04:06	18:50 ± 05:17
Semesterwoche 7	14:14 ± 04:44	26:24 ± 05:27
Semesterwoche 8	31:52 ± 06:14	18:59 ± 05:50
Semesterwoche 9	19:02 ± 05:12	31:49 ± 04:48
Semesterwoche 10	24:41 ± 06:14	16:19 ± 04:19
Semesterwoche 11	15:29 ± 04:04	30:11 ± 06:22
Ausgangstest	37:45 ± 08:05	38:59 ± 05:57
Retentionstest	38:46 ± 07:00	39:25 ± 07:01
Mittelwert gesamt:	24:19 ± 04:00	25:52 ± 04:42
Mittelwert E-Fallseminare Textformat	18:19 ± 04:17	19:35 ± 04:44
	18:56 ± 04:32	
Mittelwert E-Fallseminare Videoformat	30:21 ± 04:06	32:13 ± 04:35
	31:15 ± 04:25	

3.2.2 Ergebnisse des Eingangs-, Ausgangs- und Retentionstestes Sommersemester 2015

Zur Beantwortung der primären Forschungsfrage Nr. 1 („Wie wirkt sich das Präsentationsformat der Inhalte (Text- oder Video-*Key-Feature*) auf den Erwerb und die kurz- sowie mittelfristige Retention komplexer kognitiver Fertigkeiten (hier: Differentialdiagnostik und -therapie) aus?“) wurden die Mittelwerte der von den Studierenden erreichten Punkte in den drei formativen Prüfungen (Eingangs-, Ausgangs- und Retentionstest) in Abhängigkeit vom Präsentationsformat während der Lernphase jeweils für Text- und Videoformat miteinander verglichen.

Die Kolmogorow-Smirnow-Tests für die von den Studierenden erreichten Punkte in Ausgangs- und Retentionstest fielen signifikant aus, daher musste die Hypothese einer Normalverteilung abgelehnt werden. Die weiteren Mittelwertvergleiche wurden daher mit Wilcoxon-Tests durchgeführt.

Die Mittelwerte, jeweils für Text- und Videoformat, betrugen im Eingangstest 27,6 % ± 13,2 % und 29,4 % ± 18,3 %, im Ausgangstest 72,4 % ± 19,1 % und 76,2 % ± 19,4 %, und im Retentionstest 66,4 % ± 20,3 % und 69,2 % ± 20,2 %. Die Abbildung 7 zeigt diese Mittelwerte im zeitlichen Verlauf der Studie. Im Ausgangstest wurden von den Studierenden für Items, die in der Lernphase im Videoformat bearbeitet worden waren, signifikant mehr Punkte erreicht als für Items, die in der Lernphase im Textformat bearbeitet worden waren (Effektstärke 0,2 und $p = 0{,}039$). Im Retentionstest jedoch war dieser Unterschied zwar noch vorhanden, jedoch nicht mehr statistisch signifikant (Effektstärke 0,14).

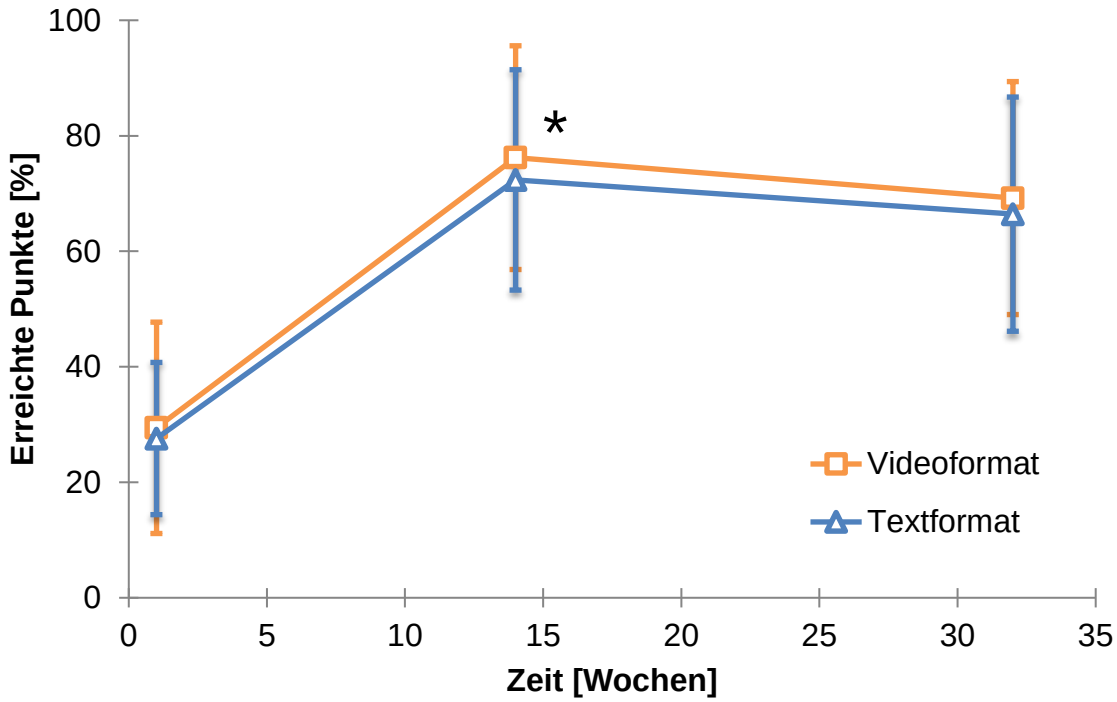


Abbildung 7: Durchschnittlich in Eingangs-, Ausgangs- und Retentionstest erreichte Punkte (Primäranalyse) in %, abhängig vom Präsentationsformat während der Lernphase. Die Balken zeigen die Standardabweichung. * $p < 0{,}05$

In den folgenden Tabellen sind die Ergebnisse für Eingangs-, Ausgangs- und Retentionstest jeweils für die Interventionsitems der Gruppe A (Tabelle 9) und die Interventionsitems der Gruppe B (Tabelle 10) dargestellt und nach einzelnen Items aufgeschlüsselt.

Tabelle 9: Interventionsitems der Gruppe A, Prozentsatz und Anzahl der Studierenden mit richtiger Antwort in den Interventionsitems der Gruppe A in Eingangs-, Ausgangs- und Retentionstest

Key Feature		**Eingangstest**			**Ausgangstest**			**Retentionstest**		
Fall	**Item**	**Gruppe A (n = 48)**	**Gruppe B (n = 45)**	**p**	**Gruppe A (n = 48)**	**Gruppe B (n = 45)**	**p**	**Gruppe A (n = 48)**	**Gruppe B (n = 45)**	**p**
1	**6**	29,2 (14)	8,9 (4)	0,012	81,3 (39)	91,1 (41)	0,174	68,8 (33)	64,4 (29)	0,664
1	**7**	14,6 (7)	13,3 (6)	0,864	89,6 (43)	93,3 (42)	0,524	93,8 (45)	97,8 (44)	0,344
1	**8**	41,7 (20)	17,8 (8)	0,011	83,3 (40)	77,8 (35)	0,503	83,3 (40)	91,1 (41)	0,264
2	**1**	75 (36)	60 (27)	0,126	91,7 (44)	91,1 (41)	0,925	87,5 (42)	88,9 (40)	0,838
2	**2**	37,5 (18)	24,4 (11)	0,176	72,9 (35)	57,8 (26)	0,127	62,5 (30)	46,7 (21)	0,128
2	**3**	20,8 (10)	31,1 (14)	0,265	66,7 (32)	71,1 (32)	0,648	70,8 (34)	68,9 (31)	0,840
2	**4**	4,2 (2)	2,2 (1)	0,601	87,5 (42)	88,9 (40)	0,838	72,9 (35)	73,3 (33)	0,964
2	**5**	50 (24)	33,3 (15)	0,105	62,5 (30)	46,7 (21)	0,128	54,2 (26)	53,3 (24)	0,937
2	**6**	18,8 (9)	15,6 (7)	0,687	58,3 (28)	62,2 (28)	0,706	50 (24)	33,3 (15)	0,105
4	**5**	64,6 (31)	68,9 (31)	0,664	85,4 (41)	88,9 (40)	0,622	81,3 (39)	91,1 (41)	0,170
4	**6**	14,6 (7)	6,7 (3)	0,218	79,2 (38)	75,6 (34)	0,681	83,3 (40)	71,1 (32)	0,165
4	**7**	27,1 (13)	8,9 (4)	0,022	72,9 (35)	62,2 (28)	0,277	77,1 (37)	62,2 (28)	0,123
4	**8**	52,1 (25)	33,3 (15)	0,069	87,5 (42)	88,9 (40)	0,838	91,7 (44)	84,4 (38)	0,291
4	**9**	60,4 (29)	44,4 (20)	0,126	64,6 (31)	33,3 (15)	0,002	45,8 (22)	33,3 (15)	0,222

Legende: *Key-Feature*-Fall und -Item beziehen sich jeweils auf Tabelle 1

Tabelle 10: Interventionsitems der Gruppe B, Prozentsatz und Anzahl der Studierenden mit richtiger Antwort in den Interventionsitems der Gruppe B in Eingangs-, Ausgangs- und Retentionstest

Key Feature		**Eingangstest**			**Ausgangstest**			**Retentionstest**		
Fall	**Item**	**Gruppe A (n = 48)**	**Gruppe B (n = 45)**	**p**	**Gruppe A (n = 48)**	**Gruppe B (n = 45)**	**p**	**Gruppe A (n = 48)**	**Gruppe B (n = 45)**	**p**
1	**1**	41,7 (20)	33,3 (15)	0,413	81,3 (39)	77,8 (35)	0,682	68,8 (33)	48,9 (22)	0,053
1	**2**	37,5 (18)	13,3 (6)	0,007	81,3 (39)	86,7 (39)	0,483	70,8 (34)	77,8 (35)	0,450
1	**3**	25 (12)	8,9 (4)	0,038	56,3 (27)	48,9 (22)	0,483	47,9 (23)	42,2 (19)	0,586
1	**4**	14,6 (7)	6,7 (3)	0,218	64,6 (31)	62,2 (28)	0,816	54,2 (26)	57,8 (26)	0,729
1	**5**	37,5 (18)	35,6 (16)	0,848	68,8 (33)	73,3 (33)	0,631	62,5 (30)	55,6 (25)	0,501
3	**1**	8,3 (4)	2,2 (1)	0,188	60,4 (29)	66,7 (30)	0,537	64,6 (31)	66,7 (30)	0,835
3	**2**	4,2 (2)	0 (0)	0,159	27,1 (13)	20 (9)	0,427	25 (12)	22,2 (10)	0,756
3	**3**	68,8 (33)	68,9 (31)	0,989	91,7 (44)	100 (45)	0,044	95,8 (46)	95,6 (43)	0,948
3	**4**	6,3 (3)	11,1 (5)	0,409	64,6 (31)	82,2 (37)	0,054	60,4 (29)	71,1 (32)	0,282
3	**5**	43,8 (21)	44,4 (20)	0,947	85,4 (41)	93,3 (42)	0,218	85,4 (41)	77,8 (35)	0,346
3	**6**	0 (0)	2,2 (1)	0,323	79,2 (38)	91,1 (41)	0,106	54,2 (26)	64,4 (29)	0,319
4	**1**	31,3 (15)	28,9 (13)	0,807	72,9 (35)	80 (36)	0,427	64,6 (31)	64,4 (29)	0,989
4	**3**	43,8 (21)	42,2 (19)	0,883	75 (36)	75,6 (34)	0,951	64,6 (31)	75,6 (34)	0,252
4	**4**	39,6 (19)	8,9 (4)	0,000	89,6 (43)	91,1 (41)	0,806	83,3 (40)	91,1 (41)	0,264

Legende: *Key-Feature*-Fall und -Item beziehen sich jeweils auf Tabelle 1

3.2.3 Analyse nach Bayes

Ergänzend zur konventionellen statistischen Auswertung wurde zur Beantwortung der primären Forschungsfrage Nr. 1 („Wie wirkt sich das Präsentationsformat der Inhalte (Text- oder Video-*Key-Feature*) auf den Erwerb und die kurz- sowie mittelfristige Retention komplexer kognitiver Fertigkeiten (hier: Differentialdiagnostik und -therapie) aus?“) zusätzlich der Bayes-Faktor berechnet.

Unter der Annahme einer Effektstärke halb so groß wie die Effektstärke der Vorstudie, in der repetitives Prüfen mit repetitivem Fallstudium verglichen wurde (Raupach et al. 2016), sowie der Differenz der Mittelwerte von 2,8 % im Retentionstest (66,4 % ± 20,3 % und 69,2 % ± 20,2 %) und dem dazugehörigen Standardfehler (1,7 %) ergibt sich ein Bayes-Faktor von 2,36. Dies unterstützt die Alternativhypothese, dass Video-*Key-Feature*-Items eine höhere mittelfristige Retention erzeugen als Text-*Key-Feature*-Items.

3.2.4 Kongruenz der Präsentationsformate

Zur Beantwortung der Forschungsfrage Nr. 2 („Welchen Einfluss hat die Kongruenz zwischen den Präsentationsformaten während der Lernphase und in den konsekutiven Prüfungen auf die kurz- sowie mittelfristige Retention?“) wurde zunächst ein Friedman-Test durchgeführt, welcher sowohl für den Ausgangstest als auch für den Retentionstest signifikant war. Im anschließend durchgeführten Wilcoxon-Vorzeichen-Rang-Test nach Bonferroni-Korrektur für multiples Testen (korrigiertes α-Niveau = 0,0083) zeigten sich die in Tabelle 11 durch hochgestellte Buchstaben gekennzeichneten signifikanten Unterschiede zwischen je zwei Werten.

Im Ausgangstest zeigten sich beispielsweise Unterschiede zwischen den Kategorien „*Lernphase: Text, Prüfungen: Text*“ und „*Lernphase: Text, Prüfungen: Video*“ ([a]), „*Lernphase: Text, Prüfungen: Video*“ und „*Lernphase: Video, Prüfungen: Text*“ ([b]) sowie „*Lernphase: Video, Prüfungen: Text*“ und „*Lernphase: Video, Prüfungen: Video*“ ([c]). Weiterhin wurden von den Studierenden stets mehr Punkte erreicht, wenn in den konsekutiven Prüfungen das Textformat verwendet wurde.

Tabelle 11: Kongruenz der Präsentationsformate zwischen Lernphase und konsekutiven Prüfungen

Präsentationsformat		**Erreichte Punkte (%) in den konsekutiven Prüfungen**	
Lernphase	**Konsekutive Prüfungen**	**Ausgangstest**	**Retentionstest**
Text	Text	77,4 ± 21,0[a]	71,6 ± 20,4[de]
Text	Video	67,3 ± 22,2[ab]	61,3 ± 25,4[df]
Video	Text	79,0 ± 20,2[bc]	72,8 ± 23,1[fg]
Video	Video	73,4 ± 22,3[c]	65,6 ± 22,5[eg]

Legende: Signifikanzen sind mit hochgestellten Buchstaben a-g gekennzeichnet

In dem Vergleich der zusammengefassten Kategorien mit kongruenten Präsentationsformaten *(„Lernphase: Text, Prüfungen: Text“* und *„Lernphase: Video, Prüfungen: Video“*) mit den inkongruenten Präsentationsformaten (*„Lernphase: Text, Prüfungen: Video“* und *„Lernphase: Video, Prüfungen: Text“*) zeigte sich kein signifikanter Unterschied. Die Nullhypothese darf folglich nicht abgelehnt werden.

3.2.5 Einfluss des Präsentationsformats auf den individuellen Lernerfolg

Zur Beantwortung von Forschungsfrage Nr. 3 („Inwiefern unterscheiden sich die Studierenden bezüglich des Geschlechts, des Alters und der Klausurvorleistung hinsichtlich der Effekte der Präsentationsformate auf ihren individuellen Lernerfolg?“) wurden jeweils für das Text- sowie das Videoformat die Studierenden je nach Wissenszuwachs vom Eingangs- zum Ausgangstest in drei gleichgroße Gruppen (je n = 31) stratifiziert. In einem nachfolgenden Schritt wurde das Drittel der Studierenden mit dem höchsten Wissenszuwachs („oberes Drittel“) mit dem Drittel der Studierenden mit dem geringsten Wissenszuwachs („unteres Drittel“) bezüglich ihres Geschlechts, ihres Alters und ihrer Klausurvorleistung verglichen.

Die Ergebnisse bezüglich der Verteilung der Geschlechter sind in Tabelle 12 zusammengefasst, es zeigte sich weder für das Text- noch das Videoformat ein signifikanter Unterschied zwischen oberem Drittel und unterem Drittel.

Tabelle 12: Einfluss des Geschlechts auf den individuellen Lernerfolg in % (n)

	Textformat		Videoformat	
	Oberes Drittel	**Unteres Drittel**	**Oberes Drittel**	**Unteres Drittel**
Männlich	29 % (9)	42 % (13)	35 % (11)	42 % (13)
Weiblich	71 % (22)	58 % (18)	65 % (20)	58 % (18)
	p = 0,288		p = 0,602	

Bei der Betrachtung des Alters zeigte sich für beide Präsentationsformate, dass das obere Drittel gegenüber dem unteren Drittel signifikant jünger war. Die Mittelwerte betrugen für das Textformat 24,7 ± 2,2 und 26,3 ± 3,5 Jahre (p = 0,041), für das Videoformat 24,7 ± 2,8 und 27,7 ± 4,5 Jahre (p = 0,003).

Bei den Klausurvorleistungen zeigte sich für das Textformat, dass Studierende des oberen Drittels im vorangegangenen Semester mit 83,75 ± 7,67 % (p = 0,048) signifikant mehr Punkte erreicht hatten als Studierende des unteren Drittels mit 79,73 ± 7,87 %. Für das Videoformat zeigte sich mit 83,33 ± 6,74 % und 81,86 ± 7,84 % kein signifikanter Unterschied (p = 0,438).

Aufgrund der gefundenen Unterschiede muss die Nullhypothese abgelehnt werden.

3.2.6 Itemanalyse und Testgütekriterien Sommersemester 2015

Zur Beantwortung von Forschungsfrage Nr. 4 („Wie unterscheiden sich die beiden Präsentationsformate Text- und Video-*Key-Feature* hinsichtlich ihrer Itemkennwerte und Testgütekriterien?“) wurden für jedes E-Fallseminar und Präsentationsformat die Itemschwierigkeiten und Trennschärfen berechnet und gemittelt. Als Maßzahl für die interne Konsistenz wurde Cronbachs α je E-Fallseminar und Präsentationsformat berechnet. Für diese Berechnungen wurden jeweils die Daten aller Studienteilnehmer, die jeweils bei einem E-Fallseminar anwesend waren, herangezogen, nicht nur die der in der Primäranalyse eingeschlossenen. Die Ergebnisse dieser Analysen sind in Tabelle 13 zusammengefasst.

Tabelle 13: Cronbachs α und gemittelte Itemkennwerte nach E-Fallseminaren (Sommersemester 2015)

E-Fallseminar Semesterwoche		Gruppe	Cronbachs α	Itemschwierigkeit	Trennschärfe	
Eingangstest		A & B	0,70	0,28 ± 0,20	0,240 ± 0,103	
2	Text	B	0,57	0,68 ± 0,29	0,229 ± 0,169	
	Video	A	0,52	0,71 ± 0,26	0,187 ± 0,137	
3	Text	A	0,49	0,62 ± 0,27	0,166 ± 0,163	
	Video	B	0,41	0,63 ± 0,32	0,133 ± 0,141	
4	Text	B	0,36	0,76 ± 0,26	0,096 ± 0,125	
	Video	A	0,37	0,77 ± 0,23	0,120 ± 0,152	
5	Text	A	0,69	0,53 ± 0,26	0,304 ± 0,101	*
	Video	B	0,31	0,55 ± 0,31	0,108 ± 0,125	
6	Text	B	0,32	0,72 ± 0,17	0,102 ± 0,068	*
	Video	A	0,56	0,77 ± 0,13	0,209 ± 0,148	
7	Text	A	0,75	0,66 ± 0,16	0,362 ± 0,110	*
	Video	B	0,57	0,61 ± 0,22	0,207 ± 0,134	
8	Text	B	0,57	0,62 ± 0,22	0,218 ± 0,115	
	Video	A	0,68	0,60 ± 0,23	0,286 ± 0,169	
9	Text	A	0,68	0,61 ± 0,17	0,292 ± 0,076	
	Video	B	0,64	0,53 ± 0,21	0,262 ± 0,174	
10	Text	B	0,70	0,71 ± 0,19	0,318 ± 0,093	*
	Video	A	0,82	0,70 ± 0,18	0,446 ± 0,095	
11	Text	A	0,75	0,71 ± 0,14	0,358 ± 0,151	*
	Video	B	0,49	0,63 ± 0,21	0,170 ± 0,130	
Ausgangstest		A & B	0,85	0,73 ± 0,15	0,397 ± 0,114	
Retentionstest		A & B	0,84	0,66 ± 0,18	0,361 ± 0,107	

Legende: Signifikante Unterschiede zw. Text- und Videoformat sind mit * gekennzeichnet

Bei acht der zehn E-Fallseminare der Lernphase war die je E-Fallseminar gemittelte Itemschwierigkeit der Video-*Key-Feature*-Items annähernd gleich (Differenz ≤ 0,05) der gemittelten Itemschwierigkeit der Text-*Key-Feature*-Items (Differenz: Mittelwert = 0,04; Minimum = 0,01; Maximum = 0,08). Bei den Trennschärfen zeigten sich trotz größerer Variabilität ebenfalls ähnliche Werte zwischen den Präsentationsformaten (Differenz: Mittelwert = 0,097; Minimum = 0,024; Maximum = 0,196).

Bei der Betrachtung der insgesamt über die Lernphase gemittelten Werte für Itemschwierigkeit (Video-*Key-Feature*-Items $0{,}65 \pm 0{,}08$ vs. Text-*Key-Feature*-Items $0{,}66 \pm 0{,}06$) und Trennschärfe (Video-*Key-Feature*-Items $0{,}21 \pm 0{,}10$ vs. Text-*Key-Feature*-Items $0{,}25 \pm 0{,}09$) sowie im Cronbachs α als Maßzahl für die interne Konsistenz (E-Fallseminare im Videoformat 0,54 vs. E-Fallseminare im Textformat 0,59) wurden keine Unterschiede zwischen den Präsentationsformaten gefunden. Folglich darf die Nullhypothese nicht abgelehnt werden.

Auf Ebene der einzelnen E-Fallseminare zeigte sich jedoch für Semesterwoche fünf, sechs, sieben, zehn und elf ein signifikanter Unterschied zwischen Text- und Videoformat in der Trennschärfe. Bei diesen E-Fallseminaren wiesen in Studiengruppe B je nach Semesterwoche sowohl Text- als auch Video-Items eine geringe Trennschärfe auf.

3.2.7 Evaluationsbogen

Zur Beantwortung von Forschungsfrage Nr. 5 („Wie beurteilen die Studierenden die unterschiedlichen Präsentationsformate Text- und Video-*Key-Feature*?“) wurde im Rahmen des letzten E-Fallseminars des Semesters eine Evaluation der E-Fallseminare durch die Studierenden durchgeführt. Es wurden 89 digitale Evaluationsbögen ausgefüllt, was einer Rückläuferquote von 84,8 % aller 105 im Ausgangstest anwesenden Studierenden entspricht. Die Evaluationsdaten wurden anonym erhoben und können nicht mit den weiteren Daten dieser Studie in Verbindung gebracht werden. Die Ergebnisse beziehen sich, soweit nicht weiter angegeben, auf eine sechsstufige Likert-Skala von eins „trifft voll zu“ bis sechs „trifft überhaupt nicht zu“ und sind in Tabelle 14 zusammengefasst.

Die Mehrheit der Studierenden stimmte der Aussage „Ich bin gerne zu den E-Fallseminaren gegangen.“ eher zu. Fast alle Studierenden halten die E-Fallseminare für eine sinnvolle Ergänzung der klinischen Lehre und würden E-Fallseminare auch für andere klinische Module und Fächer begrüßen. Der Aussage „Ich würde auch zu den E-Fallseminaren gehen, wenn diese keine Pflichtveranstaltungen wären.“ stimmte eine Mehrheit der Studierenden eher zu (Bewertung eins oder zwei), nur wenige lehnten diese eher ab (Bewertung fünf oder sechs). Auch Inhalt („Die zur Beantwortung der Frage relevanten Informationen waren verständlich in den Videos untergebracht.“), Gestaltung („Die Vi-

deos waren ansprechend gestaltet.“) und Realitätsnähe („Die Videos waren realistisch.“) der Videos erhielten Zustimmung von einer Mehrheit der Studierenden. Die Videofälle haben der Mehrheit der Studierenden (73,5 %) besser gefallen (Bewertung 1 oder 2) als die Textfälle („Die Videofälle haben mir insgesamt besser gefallen als die Textfälle.“). Daher muss die Nullhypothese abgelehnt werden. Insgesamt wurden die E-Fallseminare im Schulnotenprinzip von mehr als 90 % der Studierenden mit „sehr gut“ und „gut“ bewertet.

Tabelle 14: Evaluationsergebnisse

Aussage	1	2	3	4	5	6	Mdn	n
Allgemein								
Ich bin gerne zu den E-Fallseminaren gegangen.	49,9	29,9	10,3	4,6	4,6	1,1	2	87
Ich halte E-Fallseminare für eine sinnvolle Ergänzung der klinischen Lehre.	70,1	21,8	6,9	1,1	0	0	1	87
Ich glaube, dass ich in den E-Fallseminaren etwas gelernt habe, das für meine spätere Tätigkeit relevant ist.	65,5	26,4	3,4	4,6	0	0	1	87
Ich würde auch zu den E-Fallseminaren gehen, wenn diese keine Pflichtveranstaltungen wären.	34,5	29,9	16,1	8	4,6	6,8	2	87
Ich würde E-Fallseminare auch für andere klinische Module / Fächer begrüßen.	70,9	19,8	2,3	2,3	2,3	2,3	1	86
Die Videofälle haben mir insgesamt besser gefallen als die Textfälle.	24,1	35,6	13,8	11,5	4,6	10,3	2	87
Inhalt								
Ich habe jedes Video komplett angeschaut, bevor ich die zugehörige Frage beantwortet habe.	51,7	28,7	11,5	0	3,4	4,6	1	87
Die zur Beantwortung der Frage relevanten Informationen waren verständlich in den Videos untergebracht.	64,4	23	5,7	6,9	0	0	1	87
Die Videos waren ansprechend gestaltet.	64,4	28,7	5,7	1,1	0	0	1	87
Die Präsentation technischer Untersuchungsbefunde (Labor, EKG, Röntgen) war gut gelöst.	47,1	31	11,5	3,4	4,6	2,3	2	87
Die Videos haben mich verwirrt.	2,3	3,4	2,3	6,9	29,9	55,2	6	87
Die Videos waren zu überspitzt.	3,5	16,3	14	29,8	25,6	20,9	4	86
Die Videos waren realistisch.	16,1	40,2	32,2	9,2	2,3	0	2	87

Aussage	1	2	3	4	5	6	Mdn	n
Inhalt (fortgesetzt)								
Die Videos waren zu lang.	8	14,9	27,6	13,8	24,1	11,5	3	87
Ich hätte lieber ausführliche Anamnesegespräche und körperliche Untersuchungen gesehen und dafür längere Videos in Kauf genommen.	1,1	5,7	12,6	17,2	21,8	41,4	5	87
Feedback								
Ich habe das Feedback in der Regel gelesen.	48,3	23	14,9	9,2	3,4	1,1	2	87
Ich habe im Feedback nur nach den richtigen Antwortmöglichkeiten geschaut.	11,8	15,3	10,6	14,1	31,8	16,5	4	85
Meiner Meinung nach war das Feedback ausführlich genug.	67,8	26,4	4,6	0	1,1	0	1	87
Ich war demotiviert, wenn ich im Feedback gesehen habe, dass ich die vorangehende Frage falsch beantwortet hatte.	3,5	10,5	9,3	10,5	32,6	33,7	5	86
Ich hätte gerne genauer gewusst, warum die von mir eingegebene Antwort nicht richtig war.	7,1	15,5	15,5	11,9	27,4	22,6	4,5	84
Haben Sie das Feedback bei einem der beiden Formate (Video oder Text) mehr genutzt, als bei dem anderen?	**mehr bei Video**		**mehr bei Text**		**beide gleich**			
	9,3		11,6		79,1			86
Technische Umsetzung								
Ich war mit der technischen Umsetzung der Videos zufrieden.	55,3	24,7	5,9	9,4	3,5	1,2	1	85
Der Umgang mit diesem neuen Prüfungsformat ist mir leichtgefallen.	60	29,4	7,1	1,2	2,4	0	1	85
Abschluss								
Bitte bewerten Sie die E-Fallseminare insgesamt nach dem Schulnotenprinzip.	58,8	36,5	2,4	2,4	0	0	1	85

Legende: Bewertung der Aussagen in % von „trifft voll zu“ (1) bis „trifft überhaupt nicht zu“ (6), Mdn = Median und n = Anzahl

4 Diskussion

4.1 Wesentliche Ergebnisse

In Bezug auf die in Kapitel 1.7 ausgeführten Hypothesen konnten die folgenden Beobachtungen gemacht werden:

Hinsichtlich der ersten Hypothese, der Vermittlung von *Clinical-Reasoning*-Kompetenzen in Teilgebieten der Inneren Medizin (Kardiologie, Nephrologie, Hämatologie und Onkologie), erwies sich die Bearbeitung videobasierter *Key-Feature*-Items als effektiver für den kurzfristigen Lernerfolg als die Bearbeitung textbasierter *Key-Feature*-Items. Der Unterschied war statistisch signifikant, aber gering (Effektstärke 0,2). Im mittelfristigen Lernerfolg fand sich kein statistisch signifikanter Unterschied zwischen den beiden Präsentationsformaten. Der zusätzlich berechnete Bayes-Faktor unterstützt die Alternativhypothese, dass Video-*Key-Feature*-Items eine höhere mittelfristige Retention erzeugen als Text-*Key-Feature*-Items (vgl. Kapitel 4.2.4).

Bezüglich der zweiten Hypothese, der Kongruenz der Präsentationsformate (Text- und Videoformat) während der Lernphase und den konsekutiven Prüfungen, wurde sowohl kurz- als auch mittelfristig keine höhere Retention für kongruente Präsentationsformate (Lernphase Textformat, konsekutive Prüfungen Textformat oder Lernphase Videoformat, konsekutive Prüfungen Videoformat) gefunden. Unabhängig vom Präsentationsformat in der Lernphase erreichten die Studierenden in den konsekutiven Prüfungen für Items im Textformat signifikant mehr Punkte als für Items im Videoformat.

Für die dritte Hypothese zeigte sich bei der Betrachtung der Charakteristika der Studierenden bezüglich der Effekte auf ihren individuellen Lernerfolg, dass Studierende mit hohem Wissenszuwachs sowohl für das Text- als auch für das Videoformat signifikant jünger waren als Studierende mit geringem Wissenszuwachs. Ebenfalls hatten Studierende mit hohem Wissenszuwachs im Textformat signifikant bessere Klausurvorleistungen als Studierende mit geringem Wissenszuwachs. Für das Videoformat wurde bezüglich der Klausurvorleistungen kein Unterschied gefunden. Weder für das Text- noch für das Videoformat

wurden Unterschiede in der Geschlechterverteilung zwischen Studierenden mit hohem und geringem Wissenszuwachs gefunden.

In Bezug auf die vierte Hypothese zeigte sich, dass die Itemkennwerte und Testgütekriterien der eingesetzten Video-*Key-Feature*-Items sich insgesamt nicht signifikant von denen der bereits erprobten Text-*Key-Feature*-Items unterschieden. Das Cronbachs α für E-Fallseminare im Textformat lag im Mittel bei 0,59 und für E-Fallseminare im Videoformat im Mittel bei 0,54. Für die konsekutiven Prüfungen (Ausgangs- und Retentionstest) lag das Cronbachs α bei 0,85 und 0,84.

Hinsichtlich der fünften Hypothese zeigte sich, dass die E-Fallseminare allgemein von den Studierenden mehrheitlich positiv bewertet und als sinnvolle Ergänzung der klinischen Lehre gesehen wurden. Dabei bevorzugten mehr als die Hälfte der Studierenden Video-*Key-Feature*-Fälle gegenüber Text-*Key-Feature*-Fällen. Auch Inhalt, Gestaltung und Realitätsnähe der Videos wurde von den Studierenden positiv bewertet.

4.2 Effektivität videogestützter *Key-Feature*-Prüfungen

4.2.1 *Test-enhanced Learning*

Test-enhanced Learning ist in der medizinischen Lehre ein bekanntes Konzept und es konnte durch mehrere Studien bestätigt werden, dass wiederholtes Prüfen die Retention von klinischem Wissen und klinischen Fertigkeiten steigert. Dies gilt sowohl für grundlagenwissenschaftliche (Logan et al. 2011) als auch klinische (Larsen et al. 2009) Inhalte und konnte nicht nur für die Humanmedizin (Kromann et al. 2011), sondern auch für die Zahnmedizin (Baghdady et al. 2014) und die Pflegewissenschaften (Dobson und Linderholm 2015) nachgewiesen werden.

Die in dieser Studie im Mittel von den Studierenden korrekt gegebenen Antworten im Retentionstest nach sechs Monaten (unabhängig vom Präsentationsformat) liegen mit 67,8 % in einer vergleichbaren Größenordnung mit anderen Studien oder übersteigen diese sogar. So beantworteten in zwei verschiedenen Studien von Larsen et al. (2009; 2013a) die Probanden jeweils sechs Monate

nach einer Lehrintervention mit wiederholtem Testen lediglich 39 % und 36 % der Fragen korrekt.

Während in den meisten Studien zum *Testing Effect* unter „Laborbedingungen“ neben der eigentlichen Intervention keine weitere Exposition zum Lehrinhalt erfolgte, war die vorliegende Studie von Anfang an als *Supplemental Instruction* begleitend zum Semester geplant. Das bedeutet, dass parallel zu den E-Fallseminaren dieser Studie die regulären Lehrveranstaltungen mit Vorlesungen, Seminaren und Praktika zu jeweils kongruenten Lernzielen stattfanden und die summativen Modulklausuren in den Modulen 3.1–3.3 von den Studierenden geschrieben wurden. Hinzu kamen die individuellen Lernbemühungen der Studierenden in Vorbereitung auf die Klausuren sowie der individuelle Lernzuwachs aus Famulaturen und Praktika in der vorlesungsfreien Zeit zwischen Ausgangstestat und Retentionstest. Daraus ergaben sich zwangsläufig Störgrößen mit Einfluss auf den gemessenen Lernerfolg, die in der vorliegenden Arbeit aufgrund des Studiendesigns nicht erfasst werden konnten. Eine insgesamt eingeschränkte Vergleichbarkeit mit anderen Studien ist die Folge.

In bisherigen Studien zum *Testing Effect* wurden selten Zeiträume von mehr als drei Monaten untersucht (vgl. Kapitel 1.4). Die Ergebnisse der vorliegenden Arbeit in Bezug auf die mittelfristige Retention sind vergleichbar mit denen von Kromann et al. Diese beobachteten in einer Studie zur Fertigkeit der kardiopulmonalen Reanimation, dass *Test-enhanced Learning* nicht nur zwei Wochen nach der Intervention signifikant (Effektstärke 0,93 und $p<0,001$), sondern auch nach sechs Monaten (Effektstärke 0,4 und $p = 0,06$) mit einem höheren Lernerfolg assoziiert war, allerdings wie auch in der vorliegenden Studie nicht mehr statistisch signifikant (Kromann et al. 2009; Kromann et al. 2010).

In einer Metaanalyse von 35 Studien zu *Test-enhanced Learning* wird die mittlere Effektstärke mit 0,23 angegeben (Bangert-Drowns et al. 1991; Roediger und Karpicke 2006). Auch die Vorstudie erreichte Effektstärken (Cohen’s d) von 0,29 und 0,48 in Ausgangs- und Retentionstest (Raupach et al. 2016). Mit 0,2 für das Ausgangstestat und 0,14 für den Retentionstest liegen die Effektstärken (Cohen’s d) der vorliegenden Studie darunter. Jedoch ist die Vergleichbarkeit der Effektstärken eingeschränkt, da in der vorliegenden Studie zwei Präsentationsformate unter Verwendung von *Test-enhanced Learning* miteinander vergli-

chen wurden und nicht *Test-enhanced Learning* im Vergleich zu alternativen Lehrformaten betrachtet wurde. Allgemein kann man bei Effektstärken (Cohen's d) von 0,2 für einen kleinen, von 0,5 für einen mittleren und von 0,8 für einen großen Effekt sprechen (Cohen 1992); in der vorliegenden Studie lässt sich daher nur ein kleiner Effekt nachweisen.

Einen weiteren Einfluss auf die Ergebnisse der vorliegenden Studie könnte der von Callahan et al. beschriebene Effekt haben, dass Studierende, welche freiwillig an medizindidaktischen Studien teilnehmen, oft kompetenter sind als diejenigen, die nicht teilnehmen (Callahan et al. 2007). Dies wurde in der vorliegenden Studie nicht untersucht, jedoch dürfte dieser Effekt bei einer effektiven Teilnahmequote von 95,2 % der teilnahmeberechtigten Studierenden vernachlässigbar gering sein.

4.2.2 Einsatz von Videos in der medizinischen Lehre

Der Einsatz von audiovisuellen Lehrmaterialien in der medizinischen Lehre zur Erhöhung der Authentizität und des Lernerfolgs klingt zunächst plausibel. Untersuchungen legen nahe, dass Lehrmaterialien, welche in engerem kontextuellen Zusammenhang zur späteren Anwendung des Gelernten stehen, zu einem höheren Wissenszuwachs führen können (Godden und Baddeley 1975). Auch gibt es Hinweise in der Literatur, dass durch den Einsatz von Videos in der medizinischen Lehre tiefgreifendes Denken (engl.: „*deep critical thinking*“) erhöht werden könnte (Kamin et al. 2003; Balslev et al. 2005). Dennoch mangelt es in der Medizindidaktik an randomisierten, kontrollierten Studien mit ausreichender Stichprobengröße bezüglich der Effektivität von Videos in der medizinischen Lehre allgemein und dem Einfluss von Videos auf die Vermittlung von *Clinical-Reasoning*-Kompetenzen im Speziellen.

Vor dem Hintergrund der aktuellen Literatur lassen sich jedoch auch Bedenken bezüglich der Effektivität von Videos in der Lehre äußern. In der vorliegenden Studie bevorzugten die Studierenden das Videoformat gegenüber dem Textformat und diese Präferenz steht im Einklang mit der Literatur (vgl. Kapitel 4.6). Dem gegenüber steht die These, dass der Einsatz von Videos im Kontext von problemorientiertem Lernen tiefgreifendes Denken (engl.: „*deep critical thinking*“) behindern kann (Basu Roy und McMahon 2012). So schlussfolgerten

Basu Roy und McMahon (2012), dass die zusätzlichen visuellen Informationen möglicherweise von den medizinischen Inhalten ablenken könnten. In der vorliegenden Studie wurden die Auswirkungen der verschiedenen Präsentationsformate auf tiefgreifendes Denken nicht untersucht, jedoch ist dies eine mögliche Erklärung für den insgesamt nur geringen Effekt, der in dieser Studie beobachtet wurde.

Ähnlich dazu argumentierten auch La Rochelle et al. (2011), dass der positive Effekt einer größeren emotionalen Bindung durch eine höhere Authentizität des Präsentationsformats sowie durch die damit ebenfalls einhergehende höhere kognitive Belastung abgeschwächt oder sogar negiert werden kann. Auch Dong und Goh (2015) plädieren in ihren „*Twelve tips for the effective use of videos in medical education*“ dafür, irrelevante Informationen in Videos zu minimieren, um die kognitive Belastung zu reduzieren.

Es lässt sich nicht objektivieren, wie groß der Anteil irrelevanter Informationen in den verwendeten Videos der vorliegenden Studie war. Jedoch kann angenommen werden, dass der Anteil in den Videofällen größer war als in den entsprechenden Textfällen, da die Videos bewusst ausgestaltet wurden, sodass sie für die Studierenden ansprechend anzusehen waren. Dieses Ziel scheint erreicht worden zu sein, so stimmten nahezu alle Studierenden (98,8 %) in der begleitenden Evaluation der Aussage „Die Videos waren ansprechend gestaltet“ zu. Dennoch schienen die zur Beantwortung der *Key-Feature*-Frage relevanten Informationen verständlich in den Videos untergebracht gewesen zu sein; einer entsprechenden Aussage stimmten 93,1 % der Studierenden zu.

Es war erklärtes Ziel der vorliegenden Studie mit den Videos im Rahmen der Möglichkeiten dieses Mediums die Realitätsnähe zu erhöhen; die verwendeten Videos wurden in der vorliegenden Studie auch von einer Mehrzahl der Studierenden als realistisch bewertet. Im Vergleich dazu bewerteten die Studierenden in einer Studie von Woodham et al. (2015) die Aussagen „*The use of video brought the scenario to life.*“ und „*The use of video made the scenario more memorable.*“ lediglich neutral mit leichter Tendenz zur Zustimmung (Mittelwerte 3,49 und 3,62 auf einer 5-stufigen Likert-Skala von „*strongly disagree*“ bis „*strongly agree*“) (Woodham et al. 2015). Es darf also nicht alleine durch den Einsatz von Videos in der Lehre von einer größeren Realitätsnähe ausgegan-

gen werden, sondern sie ist auch abhängig von deren Qualität (vgl. Kapitel 4.2.3).

Eine größere Realitätsnähe kann jedoch möglicherweise zu einem größeren Anteil irrelevanter Informationen führen, was jedoch für jedes Format mit einer höheren Realitätsnähe als der des Textformats gilt, letztlich bis hin zur Interaktion mit einem realen Patienten. Es müssen zum *Clinical Reasoning* je nach Realitätsnähe zusätzliche sensorische Informationen (z. B. visuell, auditiv, olfaktorisch, taktil) verarbeitet, aktiv ein Gespräch geführt und relevante von irrelevanten Informationen unterschieden werden. Dabei ist davon auszugehen, dass gerade ein realer Patient nicht nur zur Diagnosefindung relevante Informationen äußert und präsentiert. Zu diskutieren bleibt hierbei, ob zugunsten eines potentiell höheren Lernzuwachses videobasierte Fallbeispiele inhaltlich auf die relevanten Informationen beschränkt werden sollten und inwiefern dies mit einer ansprechenden Gestaltung der Videos in Konflikt steht. Hierzu bedarf es weiterer Untersuchungen.

Bezüglich der Präsentation klinischer Befunde, zum Beispiel eines Auskultationsbefundes, bietet das Videoformat jedoch entscheidende Vorteile, da entsprechende Befunde realitätsnah in Bild und Ton wiedergegeben werden können, statt im Textformat nur beschrieben zu werden. Für Medizinstudierende ist es nicht nur wichtig, Auskultationsbefunde richtig zu interpretieren, sondern auch sicher mit der entsprechenden Terminologie umgehen zu können. Holtzman et al. (2009) untersuchten hierzu die Unterschiede der Itemkennwerte beim Einsatz von Multimedia zur Präsentation von Auskultationsbefunden im Vergleich zu der Beschreibung entsprechender Befunde im Textformat in formativen Teilen der *Step 1* (43 Items) und *Step 2 Clinical Knowledge* (CK) (51 Items) Prüfungen der *United States Medical Licensing Examination* (USMLE). Hierbei zeigte sich, dass die Multimedia-Items im Mittel eine signifikant ($p < 0{,}001$) niedrigere Itemschwierigkeit aufwiesen und damit schwerer für die Studierenden zu beantworten waren als die entsprechenden Text-Items (*Step 1*: 0,516 vs. 0,740; Step 2 CK: 0,553 vs. 0,680). Ebenfalls benötigten die Studierenden signifikant ($p < 0{,}001$) länger zur Bearbeitung der Multimedia-Items als zur Bearbeitung der Text-Items (*Step 1*: 154,2 s vs. 92,6 s; Step 2 CK: 128,3 s vs. 80,5 s). Die Trennschärfe für Multimedia-Items in der *Step 1* Prüfung war signi-

fikant ($p < 0,05$) geringer als für Text-Items (0,237 vs. 0,289). Holtzman et al. schlussfolgerten, dass es den Studierenden leichter fällt, Auskultationsbefunde im Textformat zu interpretieren als in einem authentischeren Multimedia-Format und dass weitere Untersuchungen über die Vor- und Nachteile der Nutzung von Multimedia in Prüfungen notwendig sind (Holtzman et al. 2009).

Es war nicht das Ziel der vorliegenden Studie diesen Effekt zu untersuchen. Wo möglich und angebracht wurden im Videoformat zwar multimediale Auskultationsbefunde integriert, diese jedoch zusätzlich durch gesprochene Worte analog zum Textformat interpretiert, um eine Vergleichbarkeit der beiden Präsentationsformate (Text- und Videoformat) zu gewährleisten.

Das *Key-Feature*-Item „Klinischer Verdacht auf Lungenfibrose bei Dyspnoe und inspiratorischem „*Velcro*"-Knistern" (Fall 4, *Key Feature* 7) beispielsweise enthielt im Videoformat eine Audioaufnahme des Auskultationsbefundes zusätzlich zum wörtlichen Auskultationsbefund. Die Studierenden, die dieses *Key Feature* während der Lernphase im Videoformat bearbeiteten, beantworteten dieses Item sowohl im Ausgangstest (72,9 % vs. 62,2 %) als auch im Retentionstest (77,1 % vs. 62,2 %) häufiger richtig als diejenigen Studierenden, die dieses *Key Feature* während der Lernphase im Textformat bearbeiteten. Ebenfalls wurde durch erstere Gruppe die Falschantwort „Lungenödem" weniger oft gegeben (Ausgangstest 6,3 % vs. 17,8 % und Retentionstest 2,1 % vs. 20 %). Dieser Unterschied ist möglicherweise auf eben diese in dem Video enthaltene Audioaufnahme zurückzuführen, da sensorische Informationen nicht in gleichem Maße durch Text transportiert werden können.

4.2.3 Qualität der Prüfungsmaterialien und der Implementierung

Die Qualität sowie die technische Implementierung der in der vorliegenden Studie eingesetzten Videos spielen eine zentrale Rolle, um diese für die Studierenden gleichsam lehrreich und attraktiv zum Ansehen zu machen. Dennoch erscheint eine Objektivierung oder gar Messung der Qualität der produzierten Videos schwierig. In der begleitenden Evaluation bewerteten die Studierenden zusammenfassend die Videos mehrheitlich als ansprechend gestaltet und bevorzugten die Videofälle gegenüber den Textfällen. Weiter folgten die produzierten Videos Dong und Gohs (2015) „*Twelve tips for the effective use of vi-*

deos in medical education“ (vgl. Kapitel 1.6) soweit anwendbar; hierauf soll im Folgenden näher eingegangen werden.

Durch die verwendeten Video- und Audioaufnahmegeräte, die Ausleuchtung der Szenen, die externe Audioaufnahme mittels eines Mikrofons an einer Tonangel sowie den Einsatz einer professionellen Videoschnittsoftware wurde versucht eine möglichst hohe visuelle und auditive Qualität der Videos zu erreichen (vgl. Tipp 11: *„[...], the videos should have professional quality. […] Professional quality videos contain high quality images, audio, and text.*“ (Dong und Goh 2015); Tipp 12: „*It is important to make sure that the captured video has good sound and visual quality and does not have irrelevant background noises, which can easily distract the audience.*“ (Dong und Goh 2015)). Die zur Filmproduktion notwendigen Fertigkeiten wurden vom Verfasser der vorliegenden Arbeit autodidaktisch unter der Zuhilfenahme von Lehrmaterialien erworben. Es lässt sich mutmaßen, dass durch Engagement einer professionellen Filmproduktionsfirma qualitativ noch höherwertige Videos entstanden wären, dies jedoch bei deutlich höherem finanziellem Aufwand (vgl. Tipp 12: „*One challenge is to find a balance between the cost of video production, video quality, and time commitment for production.*“ (Dong und Goh 2015)). Durch die Produktion eigener Videos konnten diese jedoch genau auf die Anforderungen dieser Studie zugeschnitten und Urheberrechtsverletzungen ausgeschlossen werden (vgl. Tipp 11: „*[...], using these videos should not violate potential copyright issues*“ (Dong und Goh 2015)). Die verwendeten Videos wurden in einer strukturierten Weise produziert. Durch das Schreiben von Drehbüchern sowie deren wiederholte Korrektur wurde die inhaltliche Kongruenz der ursprünglichen, rein schriftlichen *Key-Feature*-Fälle und der entsprechenden Drehbücher garantiert (vgl. Tipp 12: „*Scripting. A script should be written [...]*“ (Dong und Goh 2015)). Dies war essentiell, um eine Vergleichbarkeit der beiden Präsentationsformate zu gewährleisten. Eine weitere Korrekturschleife nach der Postproduktion sollte erneut die inhaltliche Kongruenz zwischen den fertigen Videos und den zugrunde liegenden schriftlichen *Key-Feature*-Fällen sowie die inhaltliche Richtigkeit sicherstellen. (vgl. Tipp 12: „*Peer review. Get another subject matter expert to review your video before publishing [...]*“ (Dong und Goh 2015)).

Während der Durchführung der E-Fallseminare im Videoformat kam es wiederholt zu technischen Störungen. Diese äußerten sich in Form von nicht mehr abspielenden Videos durch Videocodec-Fehlermeldungen oder eingefrorene Bildschirme. Diese Störungen betrafen im Wintersemester 2014/15 ca. 10 % der Studierenden in den E-Fallseminaren im Videoformat. Durch einmaligen, in seltenen Fällen auch mehrmaligen Wechsel des Computerarbeitsplatzes konnten die Studierenden die E-Fallseminare ohne Datenverlust der bisherigen Antworten fortführen. Der Einfluss dieser technischen Störungen sowie der daraus entstehenden Konsequenzen auf das Ergebnis dieser Studie konnte mit dem vorliegenden Studiendesign nicht gemessen werden. Dennoch ist ein negativer Einfluss auf den in dieser Studie gezeigten Effekt zumindest nicht auszuschließen, da gezeigt werden konnte, dass bereits kurze Unterbrechungen einer Aufgabe zu einer erhöhten Fehlerrate führen können (Altmann et al. 2014).

Obwohl es theoretisch die technische Möglichkeit gab, in den E-Fallseminaren im Videoformat in den Videos vor- und zurückzuspulen oder sich das gesamte Video erneut anzusehen, funktionierte diese Funktion nicht immer zuverlässig und resultierte regelmäßig in einer fehlerhaften Darstellung der Videos, sodass nur noch der Ton zu hören war. Dem gegenüber stehen die Texte in den E-Fallseminaren im Textformat, die beliebig oft in ihrer Gänze oder Teilen gelesen, wiederholt und nach Informationen durchsucht werden konnten. Dies wurde beispielsweise auch von Woodham et al. (2015) als Vorteil des Textformats beschrieben.

In der Natur der Sache liegend, aber auch verstärkt durch die genannten technischen Einschränkungen war es nur im Textformat möglich, den Text zu überfliegen und gezielt nach Schlüsselwörtern zur Beantwortung der *Key-Feature*-Fragen zu suchen. Dies ist eine mögliche Erklärung für die schnellere Bearbeitung der E-Fallseminare im Textformat im Vergleich zu deren Bearbeitung im Videoformat in der vorliegenden Studie (18:56 ± 04:32 vs. 31:15 ± 04:25 min:s). Auch Woodham et al. (2015) berichteten von einer längeren Dauer der POL-Lehrveranstaltung bei der Verwendung videobasierter virtueller Patienten gegenüber textbasierten. Im Vergleich zu den eingangs genannten PMPs war es einer der Vorteile von *Key Features* allgemein, dass diese kürzer sind, da eben nur die Schlüsselstellen bearbeitet werden. So können pro Zeiteinheit mehr

Key-Feature-Fälle bearbeitet werden, welches aufgrund der hohen Kontextspezifität von *Key Features* notwendig ist, um alle klinisch relevanten Fachgebiete und Situationen abzubilden (van der Vleuten und Newble 1995). Dieser Vorteil wird beim Einsatz von Video-*Key-Feature*-Fällen abgeschwächt, da diese eine längere Bearbeitungszeit benötigen. Wird die durchschnittliche Bearbeitungsdauer eines Text- und eines Video-*Key-Feature*-Items (01:16 vs. 02:05 min:s) der vorliegenden Studie zugrunde gelegt, lassen sich in einer Unterrichtsstunde von 45 Minuten im Textformat theoretisch fast 14 Items mehr bearbeiten als im Videoformat (36 vs. 22 Items). Ob dies den positiven Effekt von Video-*Key-Features* reduziert und inwiefern die *Total-Time-Hypothesis* hier ebenfalls eine Rolle spielt war nicht Gegenstand der vorliegenden Arbeit, sollte aber in zukünftigen Untersuchungen berücksichtigt werden.

4.2.4 Ökonomie

Die vorliegende Studie konnte zeigen, dass eine computerbasierte Intervention einen vorteilhaften Effekt bezüglich der Vermittlung von *Clinical-Reasoning*-Kompetenzen haben kann. Sofern eine adäquate informationstechnologische Infrastruktur zur Durchführung computerbasierter Lehrformate zur Verfügung steht, gestalten sich diese in der Durchführung weniger personal- und ressourcenintensiv als beispielsweise UaKs oder Seminare und sind flexibler einsetzbar.

So können in computerbasierten Lehrveranstaltungen verschiedenste Fallbeispiele zur Vermittlung von *Clinical-Reasoning*-Kompetenzen verwendet werden, ohne dass sich Patienten mit den zu lehrenden Krankheiten aktuell in Behandlung befinden und sich für Lehrveranstaltungen zur Verfügung stellen müssten. Dies hat den Vorteil, dass auch seltene oder akut lebensbedrohliche Erkrankungen von den Studierenden in den Fallbeispielen diagnostiziert und behandelt werden müssen und so gelehrt werden können.

In Bezug auf das Präsentationsformat für E-Fallseminare mit *Key-Feature*-Fragen zeigte sich in dieser Studie zumindest kurzfristig ein Vorteil des Videoformats gegenüber dem Textformat. Vor dem Hintergrund des Mehraufwands und deutlich höherer Kosten für die Videoproduktion muss jedoch diskutiert werden, ob die Größenordnung des gezeigten Effekts diesen Mehraufwand

rechtfertigt. Auch das Ergebnis der Analyse nach Bayes im Hinblick auf die mittelfristige Retention in der vorliegenden Studie ist mit einem Bayes-Faktor von 2,36 laut der von Wetzels et al. überarbeiteten Interpretation nach Jeffreys lediglich mit „anekdotischer Evidenz“ für die Alternativhypothese (engl.: *„anecdotal evidence for H_A*“) zu interpretieren (Jeffreys 1998; Wetzels et al. 2011). Dieser kann jedoch Informationen über die A-priori-Wahrscheinlichkeit für zukünftige Untersuchungen liefern und könnte daher hilfreich sein, zukünftige Interventionen zu evaluieren.

Bezüglich der Ökonomie von Video-*Key-Feature*-Fällen könnte es sinnvoll sein, wie Kopp et al. (2006) es bereits allgemein für (Text-) *Key-Feature*-Fälle fordert, eine nationale Datenbank auch für Video-*Key-Feature*-Fälle zu allen wesentlichen klinischen Themen zu etablieren. Nach Wissen des Verfassers der vorliegenden Arbeit existiert eine solche nationale Datenbank bislang jedoch nicht.

4.3 Kongruenz der Präsentationsformate

Aufgrund der hohen Kontextspezifität von *Clinical-Reasoning*-Kompetenzen (Eva 2005) könnte vermutet werden, dass für *Key-Feature*-Items, welche sowohl während der Lernphase als auch in den konsekutiven Prüfungen (Ausgangs- und Retentionstest) beispielsweise im Videoformat bearbeitet wurden, bei denen also eine Kongruenz zwischen Interventions- und Prüfungsformat bestand, eine höhere Retention erreicht werden würde. Vereinfacht könnte man analog dazu zu der Annahme kommen, dass ein höherer Lernerfolg erzielt werden würde, wenn eine Ähnlichkeit zwischen initialem und finalem Testformat besteht.

In der Literatur gibt es zwar Hinweise darauf, dass die Gedächtnisleistung von einer Ähnlichkeit der sog. „*cues“* (Hinweise, Aufgabenstellung) zwischen Lernphase und finalem Test abhängig sein könnte (Tulving und Thomson 1973; Larsen et al. 2013b); konträr hierzu wird jedoch auch beschrieben, dass für einen solchen Effekt bislang keine starke Evidenz gefunden werden konnte (Carpenter 2009). So zeigte eine Studie von Carpenter und DeLosh (2006) beispielsweise, dass die Leistung in einem finalen Test nicht am besten war, wenn initialer und finaler Test eine hohe Ähnlichkeit aufwiesen, sondern diese vom

Format des initialen Tests abhängig war (Carpenter und DeLosh 2006; Carpenter 2009).

In der vorliegenden Studie zeigte die Kongruenz der Präsentationsformate zwischen Lernphase und den konsekutiven formativen Prüfungen ebenfalls keinen Einfluss auf den Lernerfolg. Jedoch wurden in der vorliegenden Studie von den Studierenden in den konsekutiven Prüfungen sowohl in dem Ausgangs- als auch in dem Retentionstest *Key-Feature*-Items häufiger richtig beantwortet, welche in diesen Prüfungen im Textformat präsentiert wurden, unabhängig vom Präsentationsformat während der Lernphase. Dies steht in einem gewissen Gegensatz zur o. g. Studie von Carpenter (2009), wo ein höherer Lernerfolg vom Format des initialen Tests abhängig war. Dabei ist zu berücksichtigen, dass es sich in der vorliegenden Studie um unterschiedliche Präsentationsformate, jedoch beide im *Key-Feature*-Format, handelte, bei Carpenter hingegen um unterschiedliche Prüfungsformate. Eine Vergleichbarkeit ist daher nur eingeschränkt gegeben.

Ähnliches beobachteten auch Larsen et al. (2013b). In deren Studie zu *Test-enhanced Learning* wurden die drei Konditionen Lernen durch wiederholte Prüfungen mit standardisierten Patienten (SPs), wiederholte schriftliche Prüfungen sowie Lernen mit Notizen ohne Prüfung miteinander verglichen. Nach einer initialen Lernphase mit den drei verschiedenen Konditionen war nach sechs Monaten der primäre Endpunkt der Studie der Anteil an richtigen Antworten bei einer finalen Prüfung mit SPs; zusätzlich wurde eine Woche später eine schriftliche Prüfung durchgeführt. Hier zeigt sich, dass diejenigen Studierenden, welche in der Lernphase die Kondition wiederholte schriftliche Prüfungen hatten, in der anschließenden schriftlichen Prüfung besser abschnitten als zuvor in der finalen Prüfung mit SPs (61 % vs. 49 %). Dies begründen die Autoren mit einer von den Studierenden als einfacher empfundenen schriftlichen Prüfung im Vergleich zur Prüfung mit SPs. Als einer der Gründe hierfür wird ein *Cueing*-Effekt angeführt, wonach im Gegensatz zur Prüfung mit SPs die Fragen in der schriftlichen Prüfung Hinweise (engl. *cues*) zur Beantwortung dieser oder anderer Fragen lieferten (Larsen et al. 2013b).

Im Rahmen der summativen Prüfungen an der UMG wurden nach Wissen des Verfassers der vorliegenden Arbeit bislang ausschließlich Prüfungsfragen im

Textformat verwendet. Mit der Ausnahme von Abbildungen fand keine Verwendung multimedialer Inhalte statt. Daher wäre eine mögliche Erklärung für den in der vorliegenden Studie gezeigten Effekt, dass die Medizinstudierenden eher an Fragen im Textformat gewöhnt sind.

4.4 Einflüsse auf den individuellen Lernerfolg

In der vorliegenden Studie konnte bezüglich des Geschlechts der Studierenden weder für das Textformat noch für das Videoformat ein Effekt auf den individuellen Lernerfolg beobachtet werden. In einer Studie zu Kortisolspiegeln bei *Test-enhanced Learning* von Kromann et al. konnte jedoch gezeigt werden, dass männliche Studierende während eines Tests im Vergleich zu weiblichen Studierenden einen signifikanten Anstieg des Kortisolspiegels hatten, welcher nach zwei Wochen mit einem höheren Lernzuwachs korrelierte (Kromann et al. 2011). Ein Einfluss des Geschlechts auf die Kortisol-Ausschüttung unter Stress wurde bereits in der Literatur beschrieben (Kirschbaum und Hellhammer 1994). Eine Geschlechterspezifität des *Testing Effects* wurde jedoch auch in Übersichtsarbeiten bislang nicht beschrieben (Roediger und Karpicke 2006). So schlussfolgern auch Kromann et al., dass es unwahrscheinlich erscheint, dass in Jahrzehnten von Forschung zu *Test-enhanced Learning* ein solcher Unterschied zwischen den Geschlechtern nicht entdeckt worden wäre (Kromann et al. 2011). Jedoch war es kein Ziel der vorliegenden Studie, einen generellen Geschlechterunterschied oder den Einfluss von Kortisol und Stress bezüglich *Test-enhanced Learning* zu untersuchen; hierzu bedarf es weiterer Forschung.

Bezüglich der Klausurvorleistungen konnte für *Key-Feature*-Fälle im Textformat gezeigt werden, dass sich die Studierenden zwischen oberem und unterem Drittel des Lernerfolgs signifikant in ihren Klausurvorleistungen unterschieden. Studierende mit einem höheren Lernzuwachs erbrachten im vorangehenden Semester höhere Klausurleistungen und umgekehrt. Für das Videoformat wurde kein Unterschied festgestellt. Eine mögliche Interpretation dieses Ergebnisses wäre, dass die Klausurvorleistungen kein Prädiktor für den Lernerfolg mit *Key-Feature*-Fällen im Videoformat sind, im Gegensatz zu Fällen im Textformat. Es könnte hier mit der Kongruenz des Textformats argumentiert werden, da die Klausurvorleistungen ebenfalls in Prüfungen im Textformat erbracht wurden.

Studierende, die besser oder schlechter mit dem Textformat an sich, wie beispielsweise den summativen Modulklausuren mit Einfachauswahlfragen im Vorsemester, zurechtkamen, kamen auch entsprechend besser oder schlechter mit formativen *Key-Feature*-Fällen im Textformat zurecht. Im Vergleich mit Kapitel 4.3 bezüglich der Kongruenz der Präsentationsformate ergaben sich jedoch sowohl in der vorliegenden Studie als auch in der Literatur keine Hinweise auf einen Vorteil kongruenter Präsentationsformate. Hingegen scheinen leistungsstärkere oder -schwächere Studierende in ähnlichem Maße von *Key-Feature*-Fällen im Videoformat zu profitieren. Dies könnte als Vorteil für das Videoformat ausgelegt werden, jeden Studierenden an seinem individuellen Leistungsstand abzuholen, muss dies jedoch nicht notwendigerweise bedeuten. Dies würde eine in der Theorie möglicherweise niedrigere Trennschärfe für Video-*Key-Feature*-Items bedeuten, welche in der vorliegenden Studie jedoch nicht beobachtet wurde (vgl. Kapitel 3.2.6 und 4.5). Auch La Rochelle et al. (2012) untersuchten den Einfluss von gesteigerter Authentizität in Lehr- bzw. Prüfungsformaten auf die studentische Leistung in Abhängigkeit von der Klausurvorleistung und erzielten ähnliche Ergebnisse wie in der vorliegenden Studie. Leistungsstärkere Studierende schneiden besser nach einer Intervention im Textformat ab als leistungsschwächere; im Gegensatz zur vorliegenden Studie schienen leistungsstärkere Studierende jedoch in höherem Maße von Präsentationsformaten mit gesteigerter Authentizität zu profitieren als leistungsschwächere.

Kritisch betrachtet werden muss hier ebenfalls die in dieser Studie gewählte Methode zur Berechnung des Lernerfolgs mit der Bildung der Differenzen der erreichten Punkte zwischen Ausgangs- und Eingangstest. So werden Studierende unabhängig von ihren absoluten Leistungen in diesen beiden Tests bei gleicher Differenz in das gleiche Drittel bezüglich ihres Lernerfolgs eingeteilt. Ein Student, der im Eingangstest bereits 70 % und im Ausgangstest 90 % erreichte, hat nach dieser Methode einen geringeren Lernerfolg (20 Prozentpunkte) als ein Student, der sich von 40 % auf 70 % steigert (30 Prozentpunkte). Jedoch erscheint es auch ungleich schwieriger, jemandem, der bereits viel Wissen erworben hat, noch mehr Wissen zu vermitteln. Auch andere Studien verwendeten eine Einteilung in Drittel bezüglich studentischer (Vor-) Leistungen (La Rochelle et al. 2012). Zur Erhöhung der Trennschärfe

wurde in der vorliegenden Arbeit jedoch nur das obere und untere Drittel bezüglich des Lernerfolgs betrachtet, sodass das mittlere Drittel der Studierenden in dieser Auswertung nicht abgebildet ist.

Bezüglich des Alters zeigte sich sowohl für das Text- als auch das Video-Format, dass die Studierenden mit geringerem Wissenszuwachs („unteres Drittel“) signifikant älter waren. Es lässt sich mit den vorliegenden Daten keine Aussage dazu treffen, ob es sich bei diesen älteren Studierenden um solche handelt, die erst nach einer Wartezeit und damit später einen Studienplatz erhalten haben, die vor dem Studium noch eine Berufsausbildung absolviert oder ihre Hochschulzulassung auf dem zweiten Bildungsweg erworben haben. In Zusammenschau mit den Ergebnissen der Klausurvorleistungen, insbesondere für das Textformat, könnte die Hypothese aufgestellt werden, dass dies auf ein insgesamt niedrigeres Leistungsniveau dieser Studierenden zurückzuführen sei. Eine weitere Hypothese könnte eine allgemein geringere Affinität älterer Studierender zu Computern oder Videos sein. Auch in der Literatur wird ein möglicher Einfluss des Alters auf *Test-enhanced Learning* diskutiert, welcher durch altersabhängig unterschiedliche Strategien des Erinnerns an bereits gelernte Inhalte begründet wird (Jacoby et al. 2005; Roediger und Karpicke 2006). In der Studie von Jacoby et al. betrugen die Mittelwerte für das Alter der untersuchten Gruppen 19,6 und 75,8 Jahre und liegen damit deutlich weiter auseinander als in der vorliegenden Studie mit Differenzen von nur wenigen Jahren (Jacoby et al. 2005). Diese oben genannten Hypothesen können jedoch anhand der in dieser Studie vorliegenden Daten nicht untersucht werden. Weitere Studien sind notwendig, um diese Hypothesen bezüglich des Einflusses des Alters zu testen und die den Einflüssen zugrundeliegenden lernpsychologischen Effekte, insbesondere in Bezug auf das Videoformat, zu untersuchen.

Zusammenfassend scheint es von äußerster Wichtigkeit, bei der Entwicklung neuer Lehr- und Prüfungsformate nicht systematisch Gruppen von Studierenden zu benachteiligen. Dafür ist es notwendig sicherzustellen, dass sowohl ältere und damit möglicherweise weniger computeraffine Studierende in gleichem Maße mit einem Lehr- und Prüfungsformat umgehen können wie jüngere Studierende. Dies erfordert weitere Untersuchungen, insbesondere vor der Implementierung neuer Prüfungsformate für summative Prüfungen.

4.5 Itemkennwerte und Testgütekriterien

In der vorliegenden Studie wurde als Maß für die interne Konsistenz je E-Fallseminar beziehungsweise während der Lernphase je Präsentationsformat Cronbachs α berechnet. Dies lag für E-Fallseminare im Textformat im Mittel bei 0,59 und für E-Fallseminare im Videoformat im Mittel bei 0,54. Es gab keinen signifikanten Unterschied zwischen den Präsentationsformaten. Für den Eingangs-, den Ausgangs- und den Retentionstest lag das Cronbachs α bei 0,70, 0,85 und 0,84. In der Literatur wird ein Cronbachs α mit einem Wert von über 0,8 als gut beschrieben (George und Mallery 2003; Bortz und Döring 2006). Damit lagen die E-Fallseminare der Lernphase überwiegend darunter, was möglicherweise durch die Anzahl von nur 15 *Key-Feature*-Items je E-Fallseminar zu erklären ist. Insbesondere der Ausgangs- und der Retentionstest zeigten jedoch zufriedenstellende Werte für die interne Konsistenz. Auch die in der Vorstudie der vorliegenden Arbeit von Raupach et al. (2016) gemessenen Werte für die interne Konsistenz (0,663, 0,905 und 0,895 für den Eingangs-, den Ausgangs- und den Retentionstest) lagen in einer vergleichbaren Größenordnung (Raupach et al. 2016).

Die Itemschwierigkeiten wurden je E-Fallseminar beziehungsweise während der Lernphase je Präsentationsformat der E-Fallseminare gemittelt. Die gemessenen gemittelten Itemschwierigkeiten lagen für das Textformat zwischen 0,53 und 0,76 (Mittelwert 0,66), für das Videoformat zwischen 0,54 und 0,77 (Mittelwert 0,65). Auch hier gab es keinen signifikanten Unterschied zwischen den Präsentationsformaten. Für den Eingangs-, den Ausgangs- und den Retentionstest lagen die gemittelten Itemschwierigkeiten bei 0,28, 0,73 und 0,66. In der Literatur wird eine Itemschwierigkeit zwischen 0,2 und 0,8 (Bortz und Döring 2006) bzw. 0,4 und 0,8 (Möltner et al. 2006) empfohlen. Diese wurde im Mittel erreicht.

Die Itemtrennschärfen wurden wie die Itemschwierigkeiten je E-Fallseminar beziehungsweise während der Lernphase je Präsentationsformat der E-Fallseminare gemittelt. Die gemessenen gemittelten Itemschwierigkeiten lagen für das Textformat zwischen 0,096 und 0,362 (Mittelwert 0,245) und für das Videoformat zwischen 0,108 und 0,446 (Mittelwert 0,213). Für den Eingangs-,

den Ausgangs- und den Retentionstest lagen die gemittelten Itemtrennschärfen bei 0,240, 0,397 und 0,361. Die Itemtrennschärfe sollte je nach Quelle einen Wert von über 0,5 (Bortz und Döring 2006) bzw. über 0,3 (Möltner et al. 2006) aufweisen, um zwischen leistungsstarken und leistungsschwachen Studierenden zu differenzieren. In der vorliegenden Studie wurden diese Hürden (je nach Quelle) zwar von einzelnen Items überschritten (höchste Trennschärfe bei den Text-*Key-Feature*-Items 0,651; höchste Trennschärfe bei den Video-*Key-Feature*-Items 0,594), im Mittel jedoch nur die niedrigere der beiden Empfehlungen für Ausgangs- und Retentionstest.

Im Folgenden sollen die in der vorliegenden Studie gemessenen Itemkennwerte und Testgütekriterien orientierend mit anderen Studien verglichen werden; dies jedoch unter der Maßgabe einer eingeschränkten Vergleichbarkeit aufgrund der unterschiedlichen Charakteristika der verschiedenen Studien. Insbesondere ist zu berücksichtigen, dass in der vorliegenden Studie Itemkennwerte und Testgütekriterien von insgesamt 23 Prüfungen (zehn E-Fallseminare jeweils im Text- und Videoformat sowie Eingangs-, Ausgangs- und Retentionstest) betrachtet werden, in den folgenden zitierten Studien jedoch nur jeweils eine Prüfung.

Hatala und Norman (2002) implementierten 1998 schriftliche *Key-Feature*-Prüfungen für Studierende während einer achtwöchigen Famulatur in der Inneren Medizin mit 15 *Key-Feature*-Items. Sie berichteten ein Cronbachs α von 0,49, Itemschwierigkeit und -trennschärfe wurden nicht angegeben (Hatala und Norman 2002). In einer Studie von Fischer et al. (2005) mit 15 *Key-Feature*-Fällen und insgesamt 60 *Key-Feature*-Items ergab sich ein Cronbachs α von 0,65, Itemschwierigkeiten zwischen 0,32 und 0,83 (Mittelwert 0,56) sowie Itemtrennschärfen zwischen 0,06 und 0,56 (Mittelwert 0,27) (Fischer et al. 2005). Amini et al. (2011) untersuchten im Rahmen der zweiten *Medical Science Olympiad* im Iran verschiedene Prüfungsformen zur Beurteilung von *Clinical-Reasoning*-Kompetenzen, unter anderem 20 *Key-Feature*-Items; für diesen Teil der Prüfung ergab sich ein Cronbachs α von 0,83, Itemschwierigkeiten zwischen 0,52 und 0,79 (Mittelwert 0,68) sowie Itemtrennschärfen zwischen 0,14 und 0,539 (Mittelwert 0,43) (Amini et al. 2011). Damit liegen die in der vorliegenden Studie erreichten Werte für die interne Konsistenz der einzelnen E-

Fallseminare sowie die gemittelten Itemkennwerte in einem Bereich, der vergleichbar ist mit anderen Studien, die *Key-Feature*-Fälle untersuchten.

Zusammenfassend lassen sich in der vorliegenden Studie keine signifikanten Unterschiede hinsichtlich der Itemkennwerte und Testgütekriterien zwischen *Key-Feature*-Prüfungen im Text- und Videoformat feststellen. Folglich darf die Nullhypothese zu Forschungsfrage vier nicht abgelehnt werden.

4.6 Rezeption durch die Studierenden

In mehreren Studien bevorzugten Studierende im Kontext von POL als Präsentationsformat von Fallbeispielen das Videoformat gegenüber dem Textformat. Dies konnte auch in der vorliegenden Studie beobachtet werden. Die Aussage „Die Videofälle haben mir insgesamt besser gefallen als die Textfälle.“ wurde hier von 73,5 % der Studierenden mit einer eins, zwei oder drei auf einer sechsstufigen Likert-Skala von eins „trifft voll zu“ bis sechs „trifft überhaupt nicht zu“ bewertet. Dies ist vergleichbar mit den Ergebnissen einer Studie von Chan et al. (2010), in welcher 75,5 % der Befragten die Aussagen „*I prefer to use PBL video triggers rather than PBL paper cases at PBL tutorials.*“ mit einer vier, fünf oder sechs ebenfalls auf einer sechsstufigen Likert-Skala von eins „*strongly disagree*“ bis sechs „*strongly agree*“ bewerteten. In einer Studie von Basu Roy und McMahon (2012) bevorzugten ebenfalls 60 % der Studierenden und 74,6 % der Tutoren das Videoformat („*prefer video*“ oder „*strongly prefer video*“) gegenüber dem Textformat. Dennoch gibt es auch Studien, in denen die Mehrheit der Studierenden (65 % vs. 35 %) das Textformat gegenüber dem Videoformat bevorzugten (Woodham et al. 2015). Gleichzeitig kritisierten die Studierenden in der Studie von Woodham et al. (2015) die Videos aufgrund schlechter Audioqualität und dadurch geringer Sprachverständlichkeit, zu vieler Einstellungswechsel und schlechter schauspielerischer Leistung. Es ist daher anzunehmen, dass diese Kritikpunkte dort mögliche Gründe für das Bevorzugen des Textformates darstellten und somit einen Hinweis auf den Einfluss der Qualität der verwendeten Videos (vgl. Kapitel 4.2.3) auch auf die Rezeption durch die Studierenden liefern.

4.7 Stärken und Limitationen der Studie

Die vorliegende Studie war nach Wissen des Verfassers der vorliegenden Arbeit die erste Studie, welche *Key-Feature*-Fälle im Videoformat einsetzte, um mittels *Test-enhanced Learning Clinical-Reasoning*-Kompetenzen zu vermitteln. Die in dieser Studie verwendeten Fälle aus Teilgebieten der Inneren Medizin beziehen sich dabei auf klinische Probleme von hoher Prävalenz oder Wichtigkeit und wurden durch die Vorstudien umfangreich pilotiert und optimiert (Raupach et al. 2016). Aufgrund der Kontextspezifität bezüglich situationsbedingter und persönlicher Faktoren, welcher *Clinical-Reasoning*-Prozesse unterliegen (Eva 2005), und da die vorliegende Studie monozentrisch an der UMG durchgeführt wurde sowie thematisch nur Teilgebiete der Inneren Medizin abdeckte, ist eine Generalisierbarkeit der Forschungsergebnisse auf andere Fachgebiete sowie Fakultäten nicht gegeben. Somit können keine sicheren Aussagen über das Potential von *Test-enhanced Learning* unter der Verwendung von *Key-Feature*-Fällen zum Erwerb von *Clinical-Reasoning*-Kompetenzen für andere Fachgebiete getroffen werden.

Durch eine stratifizierte Randomisierung der Studierenden nach Geschlecht und Klausurvorleistungen sowie die Vorgabe des Präsentationsformats für jede Gruppe und jedes E-Fallseminar sollte eine systematische Verzerrung der beiden Studiengruppen minimiert und so die interne Validität erhöht werden. Auch wenn mit 95,2 % (119 von 125) ein Großteil der berechtigten Studierenden einwilligte, an der Studie teilzunehmen, handelt es sich dennoch um eine kleine Stichprobengröße, die gegebenenfalls nicht ausreicht, um kleinere Effekte zu detektieren.

Nach dem Axiom „*assessment drives learning*“ (vgl. Kapitel 1.3.1) generieren summative Prüfungen einen größeren Lernanreiz und -erfolg als formative Prüfungen (Raupach et al. 2013). Da in der vorliegenden Studie ausschließlich formative Prüfungen verwendet wurden und auch die leistungsunabhängige Präsentation von Einfachauswahlfragen aus vorangehenden Modulklausuren vermutlich nur ein geringer zusätzlicher Anreiz für die Studierenden war, könnte davon ausgegangen werden, dass der gezeigte Effekt unter der Verwendung summativer *Key-Feature*-Prüfungen womöglich größer gewesen wäre. Dies

sollte in zukünftigen Untersuchungen berücksichtigt werden, beispielsweise durch die leistungsabhängige Vergabe von zusätzlichen Modulpunkten oder anderen Anreizen.

In der vorliegenden Studie wurde keine Adjustierung der Ergebnisse nach der Bearbeitungszeit vorgenommen, obwohl die Bearbeitung der E-Fallseminare im Videoformat im Mittel über beide Gruppen rund zwölf Minuten länger dauerte als die Bearbeitung der E-Fallseminare im Textformat. Es war jedoch kein Ziel der vorliegenden Studie den *Testing Effect* weiter bezüglich der *Total-Time-Hypothesis* und der *Retrieval-Hypothesis* zu untersuchen (vgl. Kapitel 1.4).

4.8 Ausblick und Implikationen für die weitere Forschung

In der vorliegenden Studie wurde nicht untersucht, welchen Einfluss die persönliche Präferenz der Studierenden bezüglich des Präsentationsformats, also ob das Text- oder Videoformat zur Bearbeitung von *Key-Feature*-Fällen bevorzugt wird, auf die kurz- sowie mittelfristige Retention des Gelernten hat. In einer Folgestudie von Schuelper et al. (2019), in welcher die Studierenden für jedes E-Fallseminar frei wählen konnten, ob sie dieses im Text- oder Videoformat bearbeiten wollten, erreichten Studierende, welche in den E-Fallseminaren überwiegend das Videoformat wählten, im Ausgangstestat höhere Ergebnisse als diejenigen, die überwiegend das Textformat wählten (76,2 % vs. 70,0 % p = 0,15). Im Retentionstest nach weiteren vier Monaten erreichte erstere Gruppe sogar signifikant höhere Ergebnisse (75,3 % vs. 63,4 % p = 0,02) (Schuelper et al. 2019).

Nach Cook et al. (2019) kann *Clinical Reasoning* in zwei Teilbereiche unterschieden werden, *Diagnostic Reasoning* und *Management Reasoning*. Ersteres hat das Ziel, aus den präsentierten Symptomen sowie Ergebnissen von Untersuchungen und Tests eine Diagnose zu stellen, während sich Letzteres mit Entscheidungen über Behandlung, weitere Diagnostik, Folgetermine, Nachsorge und Nutzung von Ressourcen als deutlich komplexer darstellt. Während sich der Großteil der bisherigen Forschung und medizinischen Lehre auf *Diagnostic Reasoning* fokussierte, sei relativ wenig über *Management Reasoning* und dessen kognitive Prozesse bekannt. Es war kein Ziel der vorliegenden Studie zwi-

schen *Diagnostic Reasoning* und *Management Reasoning* zu unterscheiden, dennoch zielten die verwendeten *Key-Feature*-Fragen auf beide Teilbereiche von *Clinical Reasoning* ab. Weitere Untersuchungen sind notwendig, um zu erörtern, ob sich videogestützte *Key-Feature*-Prüfungen im gleichen Maße für *Diagnostic* und *Management Reasoning* eignen. Möglicherweise bieten sich videogestützte *Key-Feature*-Prüfungen als weitergehende Lehr-, Beurteilungs- und damit Forschungsmethode für weitere Untersuchungen bezüglich von *Management-Reasoning*-Prozessen an (Cook et al. 2019).

Unklar ist ebenfalls, ob wiederholtes Prüfen mit *Key-Feature*-Fragen auch außerhalb der Lehr- und Lernumgebung eine Auswirkung auf die *Clinical-Reasoning*-Kompetenz der Studierenden in der klinischen Praxis hat. Ein solcher Zusammenhang läge nahe und wurde bereits von anderen Forschern postuliert (Larsen et al. 2013b). Jedoch sind hierfür weitere Untersuchungen notwendig, zum Beispiel um herauszufinden, ob Mediziner, die durch wiederholtes Prüfen mit *Key-Feature*-Fällen *Clinical-Reasoning*-Kompetenz trainiert haben, in der klinischen Umgebung bessere Behandlungsergebnisse erzielen und welchen Einfluss das Präsentationsformat darauf hat.

Nachdem durch mehrere Studien die Effektivität „herkömmlicher“ Text-*Key-Feature*-Fälle in der klinischen Lehre gezeigt werden konnte, hielten diese nach und nach nicht nur Einzug in die Curricula verschiedener Fakultäten, sondern im Jahr 2018 auch in das deutsche Staatsexamen (Vogel 2018). Solang dieses jedoch noch in schriftlicher Papierform abgehalten wird sind die technischen Voraussetzungen zur Implementierung von Video-*Key-Feature*-Prüfungen nicht gegeben. Bis dieses soweit ist müssen noch viele weitere offene Fragen bezüglich der Effektivität videogestützter *Key-Feature*-Prüfungen beantwortet werden.

5 Zusammenfassung

Die Vermittlung differentialdiagnostischer und -therapeutischer Kompetenzen als Teil von *Clinical Reasoning* hat einen zunehmend hohen Stellenwert in der Ausbildung zukünftiger Ärztinnen und Ärzte. Zur Vermittlung dieser Kompetenzen stehen unterschiedliche Lehr- und Prüfungsformate zur Verfügung. Im Gegensatz zu den bislang im Medizinstudium in Deutschland vorherrschenden Einfachauswahlfragen legt die Literatur nahe, dass dieses prozedurale Wissen jedoch besser durch das *Key-Feature*-Format geprüft, aber auch gelehrt werden kann. In einer Vorstudie konnte gezeigt werden, dass das wiederholte Bearbeiten von *Key-Feature*-Fällen im Textformat im Vergleich zum wiederholten Lesen derselben Fallbeispiele zu einer höheren Retention des Gelernten führte. Zudem eignen sich Videos dazu, die Authentizität von medizinischen Fallbeispielen gegenüber dem Textformat zu erhöhen. Bislang lagen keine Daten bezüglich der Effektivität videogestützter *Key-Feature*-Prüfungen beim Erwerb differentialdiagnostischer und therapeutischer Kompetenzen im Medizinstudium vor. Die vorliegende Studie ging dieser Fragestellung nach.

Dazu erfolgte die Erstellung entsprechender Video-*Key-Feature*-Fälle zu inhaltlich kongruenten, in Vorstudien bereits erprobten Text-*Key-Feature*-Fällen. Diese wurden in Form computerbasierter, elektronischer Fallseminare an der Universitätsmedizin Göttingen in der curricularen Lehre im dritten klinischen Semester im Wintersemester 2014/15 implementiert und pilotiert sowie die eigentliche Studie im Sommersemester 2015 durchgeführt. Die Studierenden bearbeiteten in diesem Cross-over-Studiendesign in den elektronischen Fallseminaren nach einem initialen Eingangstest in der darauffolgenden Lernphase im wöchentlichen Wechsel *Key-Feature*-Fälle im Text- oder im Videoformat. Es folgten zwei konsekutive Prüfungen, ein Ausgangstest in Semesterwoche zwölf und ein Retentionstest nach einem halben Jahr. Alle elektronischen Fallseminare hatten formativen Charakter, waren daher nicht benotet. Der primäre Endpunkt der Studie war der Unterschied der Mittelwerte der von den Studierenden erreichten Punkte im Ausgangstest in Abhängigkeit von dem Präsentationsformat eines Items (Text- oder Videoformat) während der Lernphase als Maß für den kurzfristigen Lernerfolg. Zusätzlich wurden die Studierenden in einem Evalua-

tionsbogen nach dem Ausgangstest bezüglich der E-Fallseminare und Video-*Key-Feature*-Fälle befragt. Es wurden die Daten von 93 Studierenden (effektive Rücklaufquote 74,4 %) in die Auswertung eingeschlossen. In der Analyse des primären Endpunktes zeigte sich, dass die Studierenden in Items, welche in der Lernphase im Videoformat bearbeitet worden waren, signifikant mehr Punkte erreichten als für Items, die in der Lernphase im Textformat bearbeitet worden waren (76,2 % ± 19,4 % vs. 72,4 % ± 19,1 % Effektstärke 0,2 und $p = 0{,}039$). Der zusätzlich berechnete Bayes-Faktor (2,36) unterstützt ebenfalls die Alternativhypothese, dass Video-*Key-Feature*-Items eine höhere mittelfristige Retention erzeugen als Text-*Key-Feature*-Items.

Die Itemkennwerte (Itemschwierigkeit und Itemtrennschärfe) der eingesetzten Video-*Key-Feature*-Items unterschieden sich insgesamt nicht signifikant von denen der bereits in Vorstudien erprobten Text-*Key-Feature*-Items und lagen in einem akzeptablen Bereich. Auch für das Cronbachs α als Maßzahl für die interne Konsistenz zur Abschätzung der Reliabilität der E-Fallseminare gab es insgesamt keinen signifikanten Unterschied zwischen den E-Fallseminaren im Textformat und denen im Videoformat. Für den Ausgangstest (primärer Endpunkt) lag das Cronbachs α bei 0,85.

Die Mehrheit der Studierenden halten die E-Fallseminare für eine sinnvolle Ergänzung der klinischen Lehre und sind der Meinung, dass sie darin etwas gelernt haben, das für ihre spätere Tätigkeit relevant sei. Sie würden E-Fallseminare auch für andere klinische Module und Fächer begrüßen. Insgesamt haben die *Key-Feature*-Fälle im Videoformat der Mehrheit der Studierenden besser gefallen als die *Key-Feature*-Fälle im Textformat. In sechs Schulnoten („sehr gut“–„ungenügend“) wurden die elektronischen Fallseminare in der Form der vorliegenden Studie von mehr als 90 % der Studierenden mit „sehr gut“ und „gut“ bewertet. Die Ergebnisse der vorliegenden Arbeit liefern Hinweise darauf, dass videobasierte *Key-Feature*-Prüfungen ein effektiver Ansatz zur Vermittlung von differentialdiagnostischen und -therapeutischen Kompetenzen sein können und Akzeptanz unter den Studierenden finden. Zukünftige Studien sollten unter anderem untersuchen, wie videobasierte *Key-Feature*-Prüfungen effektiv im Medizinstudium eingesetzt werden können und inwiefern ein Transfer dieses Effekts auf die klinische Tätigkeit in der Realität gelingen kann.

6 Anhang

Die in dieser Arbeit verwendeten Prüfungsmaterialien, also die Text- und Video-*Key-Feature*-Fälle sowie -Items, sind urheberrechtlich geschützt, können aber auf schriftliche Anfrage beim Verfasser dieser Arbeit eingesehen werden.

E-Mail: dissertation-ludwig@gmx.de

Tabelle A1: Übersicht über die verwendete Software

Software	**Entwickler**	**Ort**
CAMPUS-Prüfungssystem (Version 1.3)	Zentrum für virtuelle Patienten, Medizinische Fakultät Heidelberg	Heidelberg, Deutschland
Celtx (Version 2.9.7)	Grayfirst Corporation	St. John's, Kanada
EvaSys (Version unbekannt)	Electric Paper Evaluationssysteme GmbH	Lüneburg, Deutschland
Final Cut Pro (Versionen 10.1.2–10.2.1)	Apple Inc.	Cupertino, Vereinigte Staaten von Amerika
HandBrake (Versionen 0.10 Beta 1–0.10.2)	The HandBrake Team	ohne Ortsangabe
Item Management System (Version unbekannt)	Umbrella Consortium for Assessment Networks	Heidelberg, Deutschland
Microsoft Windows Media Player (Version 12)	Microsoft Corporation	Redmond, Vereinigte Staaten von Amerika
NetMan (Version unbekannt)	H+H Software GmbH	Göttingen, Deutschland
SPSS Statistics (Version 22)	IBM Inc.	Chicago, Vereinigte Staaten von Amerika

Tabelle A2: Übersicht über das verwendete Material für Video- und Audioaufnahmen

	Modell	**Hersteller**	**Ort**
Videotechnik			
Videokamera	HC-X810 HC-X929 HC-V500	Panasonic Corporation	Kadoma, Japan
	S110 EOS 70D	Canon Inc.	Tokio, Japan
Tontechnik			
Audiorekorder	DN-F20R Portabel IC Recorder	Denon Corporation	Kawasaki, Japan
	H4nSP	Zoom Corporation	Tokio, Japan
Mikrofon Mikrofonspeisung	ME66 K6	Sennheiser electronic GmbH & Co. KG	Wedemark, Deutschland
Lichttechnik			
Scheinwerfer	Ianebeam 800 W Nr. 3142	Ianiro S.r.l.	Trento, Italien

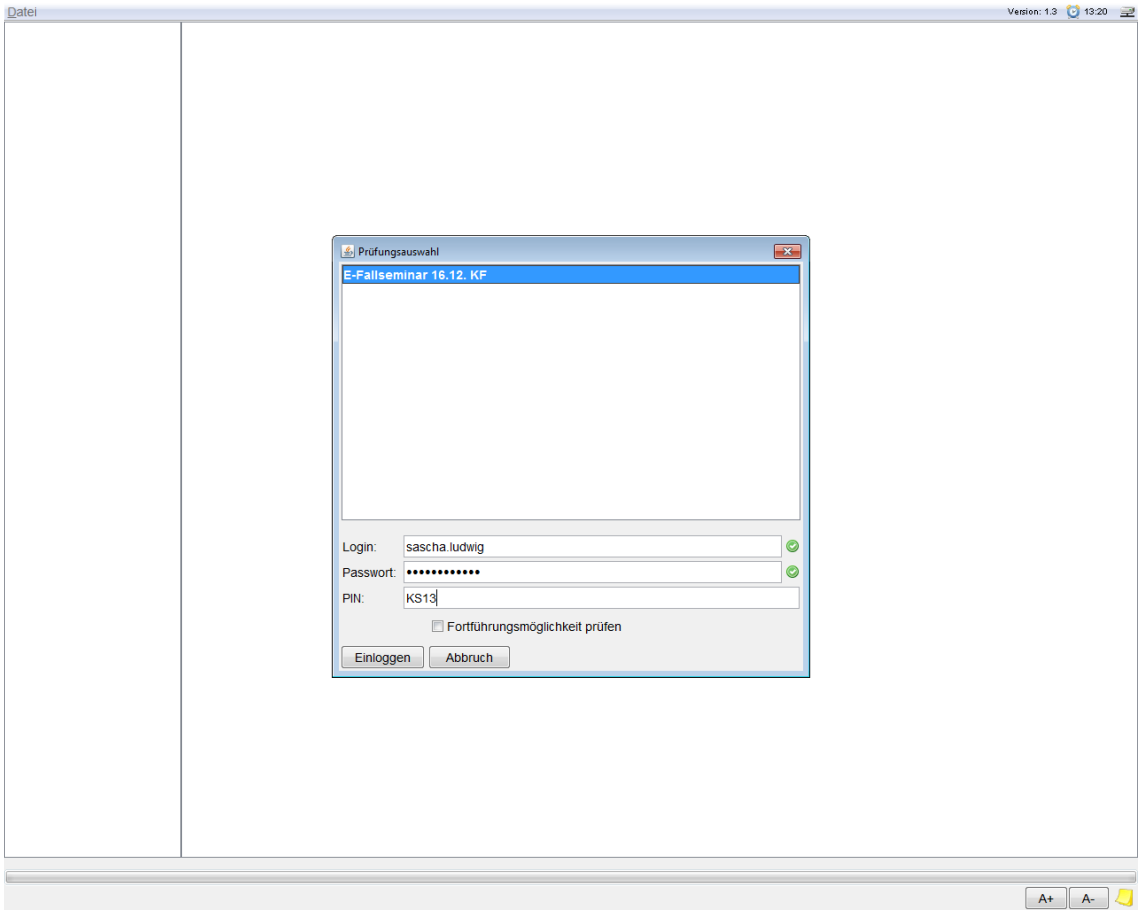


Abbildung A1: Bildschirmfoto der Anmeldemaske im CAMPUS-Prüfungsplayer mit bereits eingegebenem Benutzernamen, Passwort und PIN.

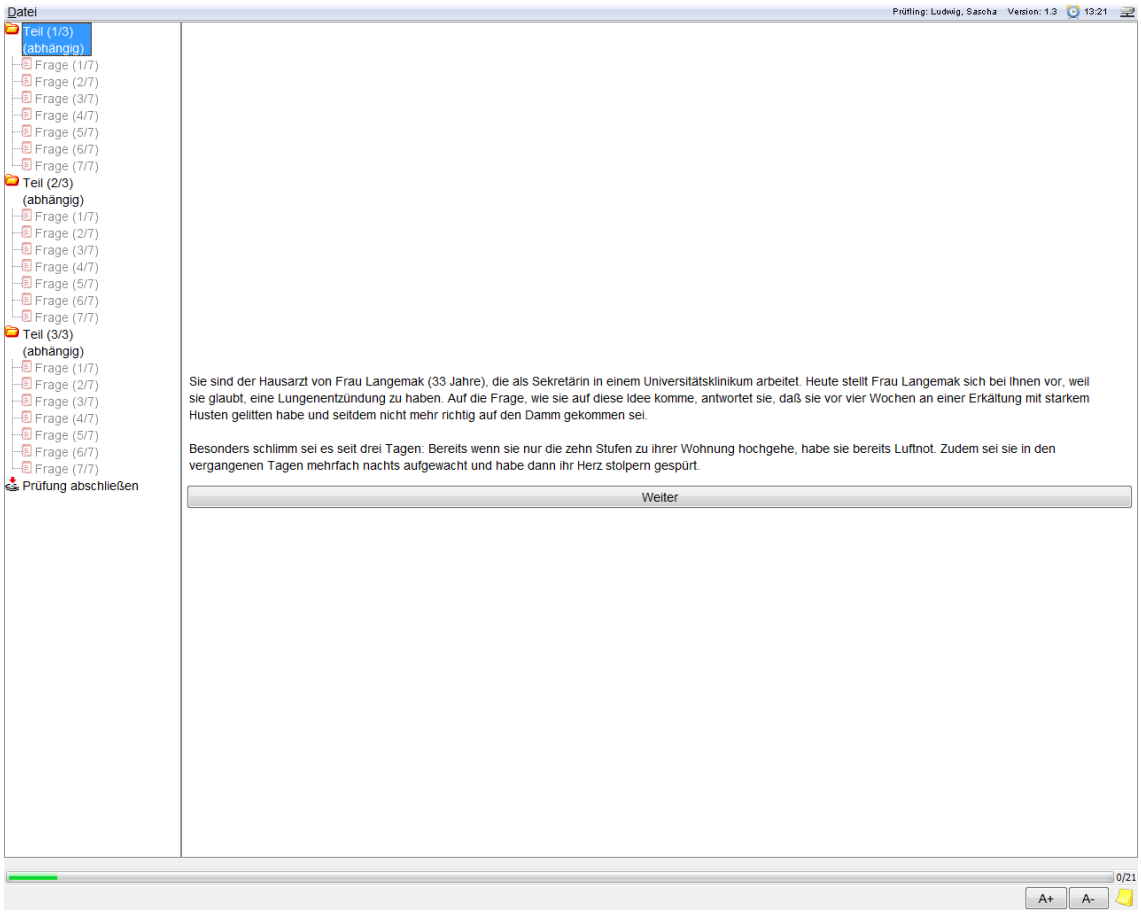


Abbildung A2: Bildschirmfoto eines Text-*Key-Feature*-Falls. *CAMPUS-Prüfungsplayer* mit Fallvignette. Durch einen Klick auf die Schaltfläche „Weiter“ gelangen die Studierenden zur ersten *Key-Feature*-Frage.

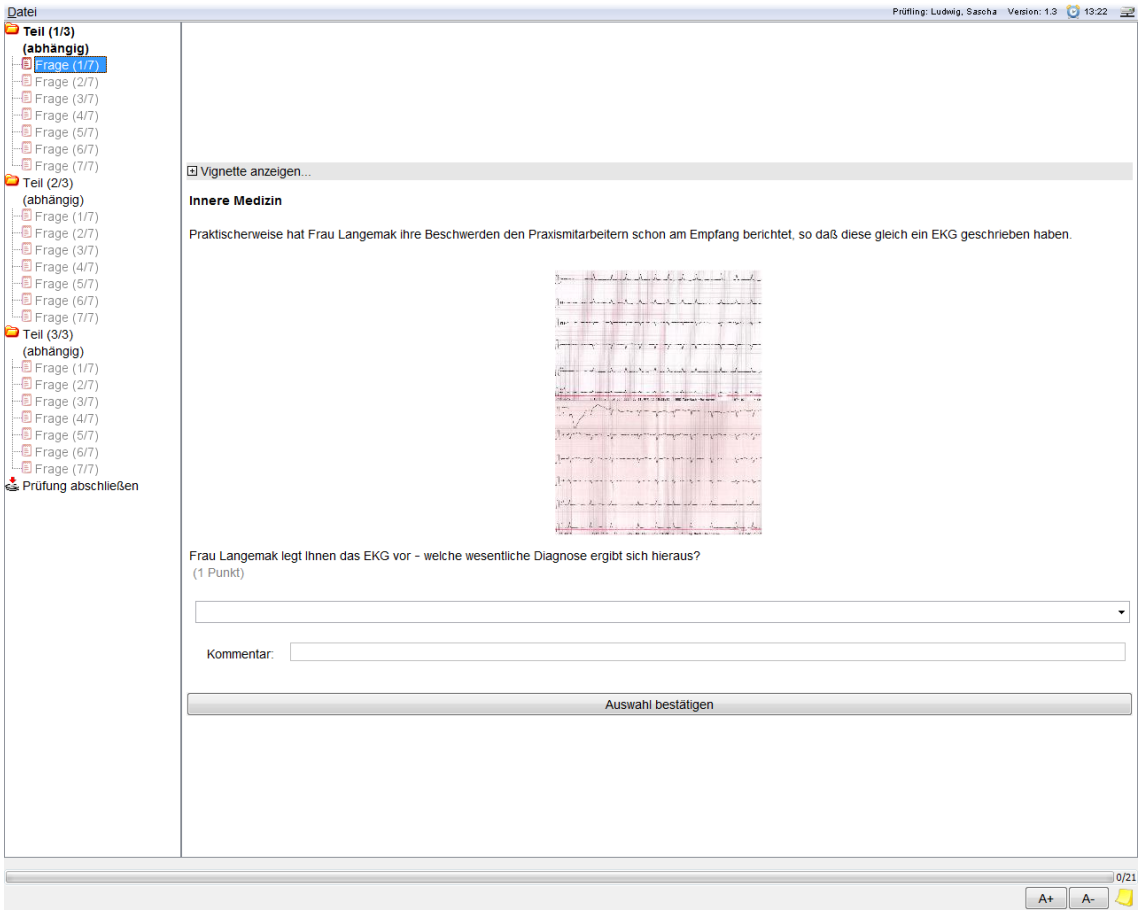


Abbildung A3: Bildschirmfoto eines Text-*Key-Feature*-Falls. *CAMPUS-Prüfungsplayer* mit Darstellung des ersten Text-*Key-Feature*-Items. Bei Bedarf kann durch Klick auf die „+"-Schaltfläche neben dem Text „Vignette anzeigen…" die Fallvignette erneut eingeblendet werden. Es folgt ein kurzer, fortführender Text und darauf die Darstellung eines Miniaturbildes eines für den Fall relevanten EKGs. Durch einen Klick auf das Miniaturbild öffnet sich das Bild des EKGs vergrößert in einem neuen Fenster (vgl. Abbildung A4). Darunter befindet sich die eigentliche *Key-Feature*-Frage sowie das noch leere Eingabefeld für die Antwort im *Long-Menu*-Format. Darunter befindet sich ein Eingabefeld mit der Möglichkeit, einen Kommentar zu verfassen. Nach Eingabe einer Antwort in das Eingabefeld kann durch Klick auf die Schaltfläche „Auswahl bestätigen" die Auswahl bestätigt werden und die Studierenden gelangen zum nächsten *Key-Feature*-Item.

Abbildung A4: Bildschirmfoto eines Text-*Key-Feature*-Falls. *CAMPUS-Prüfungsplayer* mit vergrößerter Darstellung des EKGs in einem neuen Fenster. Durch Klick auf die Schaltfläche „Schließen“ schließt sich das Fenster wieder und die Studierenden gelangen zurück zur vorherigen Ansicht.

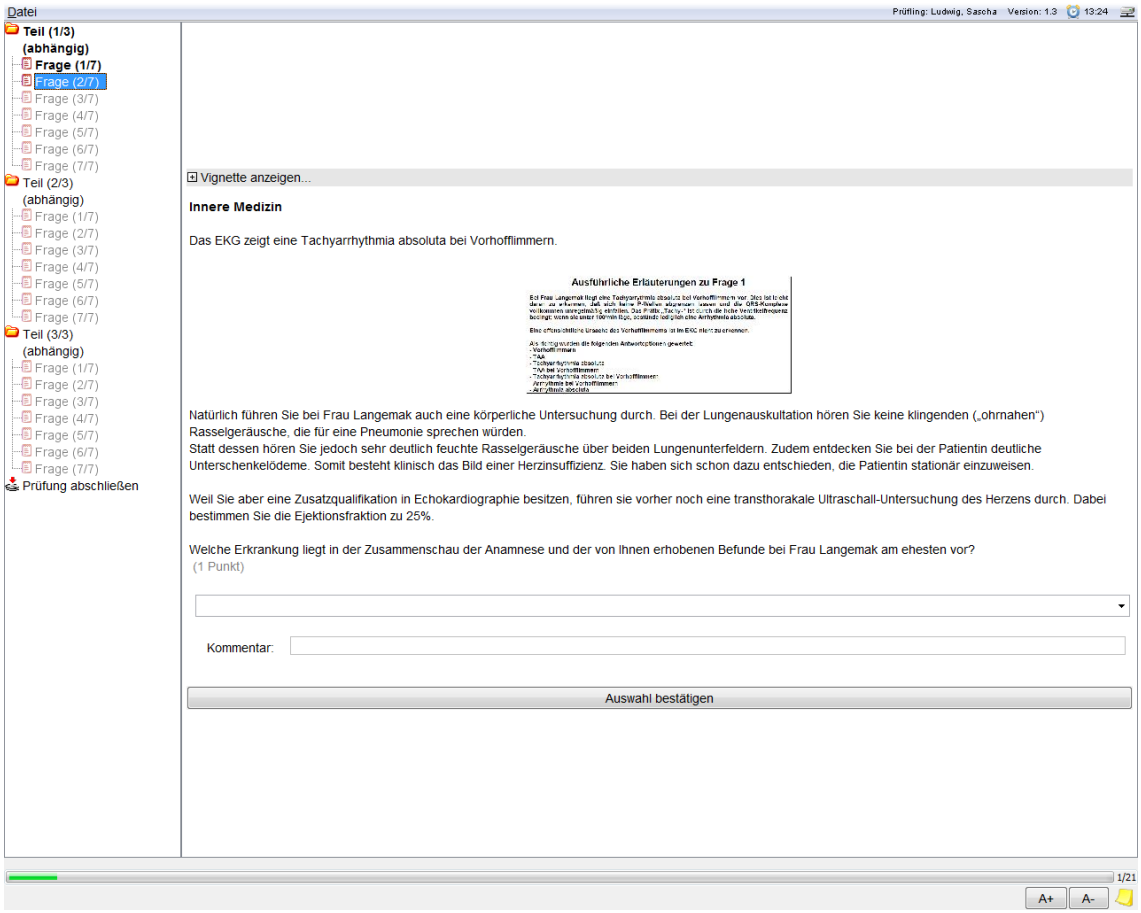


Abbildung A5: Bildschirmfoto eines Text-*Key-Feature*-Falls. *CAMPUS-Prüfungsplayer* mit der Darstellung des zweiten Text-*Key-Feature*-Items. Es folgt ein kurzer Text mit der Antwort auf die vorherige *Key-Feature*-Frage gefolgt von einem Miniaturbild von ausführlichen Erläuterungen zu Frage eins (vgl. Abbildung A6), welches sich analog zum EKG durch Klick auf selbiges vergrößern lässt. Es folgt ein Text mit der Fortführung des *Key-Feature*-Falles und der zweiten *Key-Feature*-Frage.

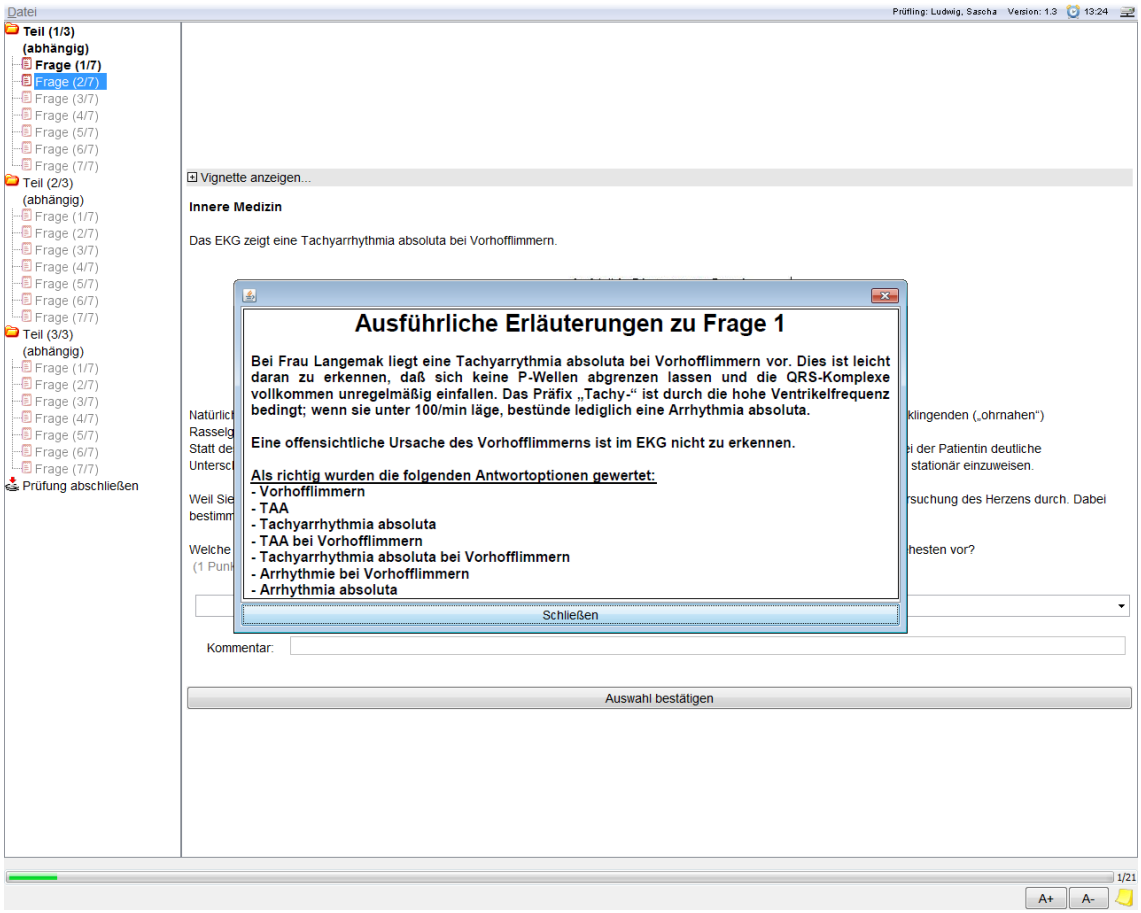


Abbildung A6: Bildschirmfoto eines Text-*Key-Feature*-Falls. *CAMPUS-Prüfungsplayer* mit der vergrößerten Darstellung der ausführlichen Erläuterungen zu Frage eins sowie einer Liste mit den als richtig gewerteten Antwortoptionen. Durch Klick auf die Schaltfläche „Schließen“ schließt sich das Fenster wieder und die Studierenden gelangen zurück zur vorherigen Ansicht.

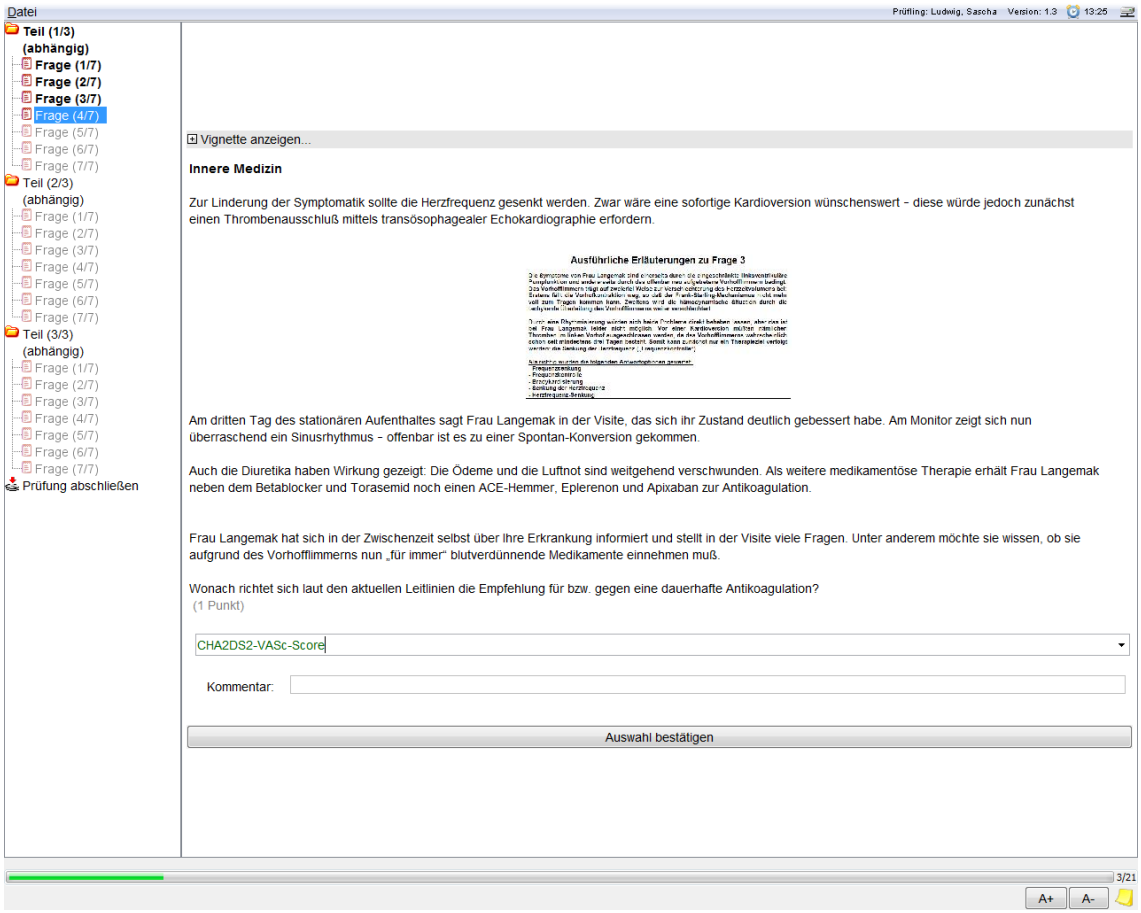


Abbildung A7: Bildschirmfoto eines Text-*Key-Feature*-Falls. *CAMPUS-Prüfungsplayer* mit der Darstellung des vierten Text-*Key-Feature*-Items und bereits eingegebener Antwort im Eingabefeld.

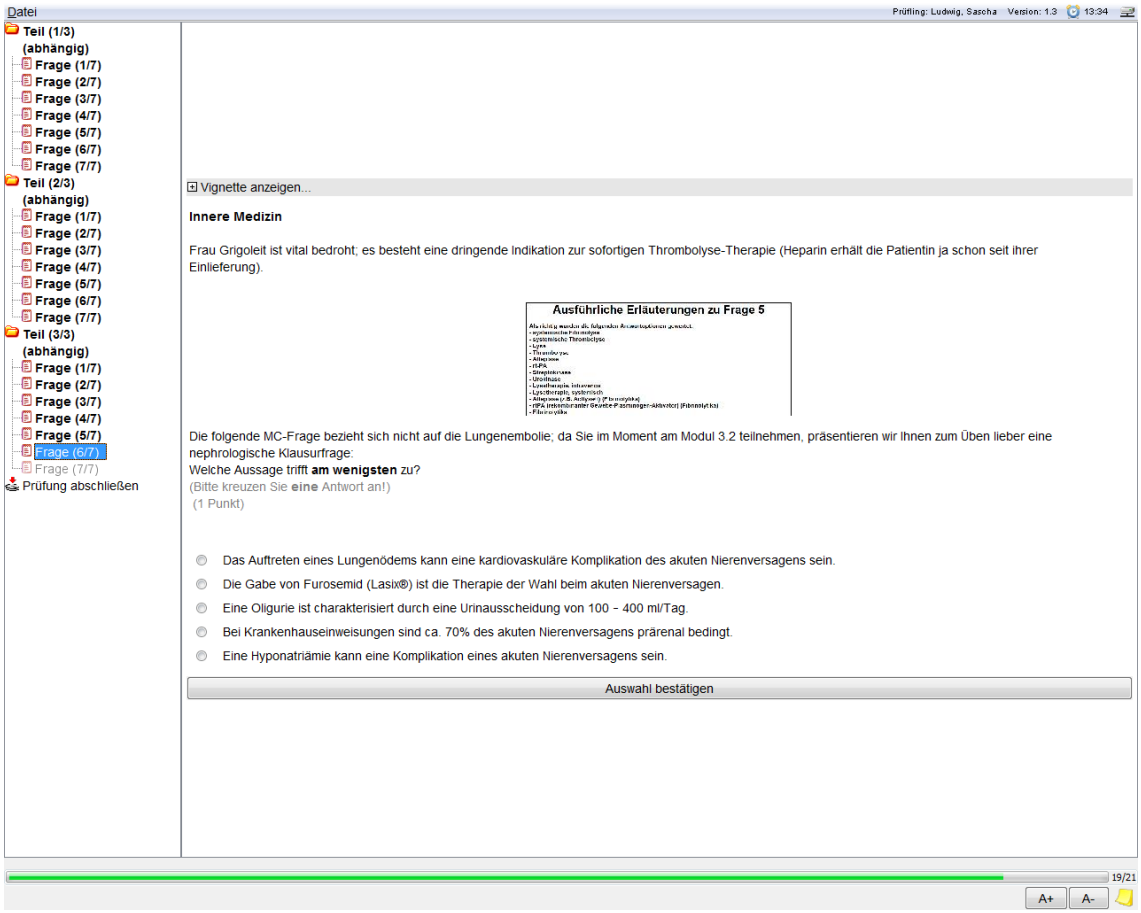


Abbildung A8: Bildschirmfoto eines Text-*Key-Feature*-Falls. *CAMPUS-Prüfungsplayer* mit der Darstellung einer Einfachauswahlfrage nach einem Text-*Key-Feature*-Fall.

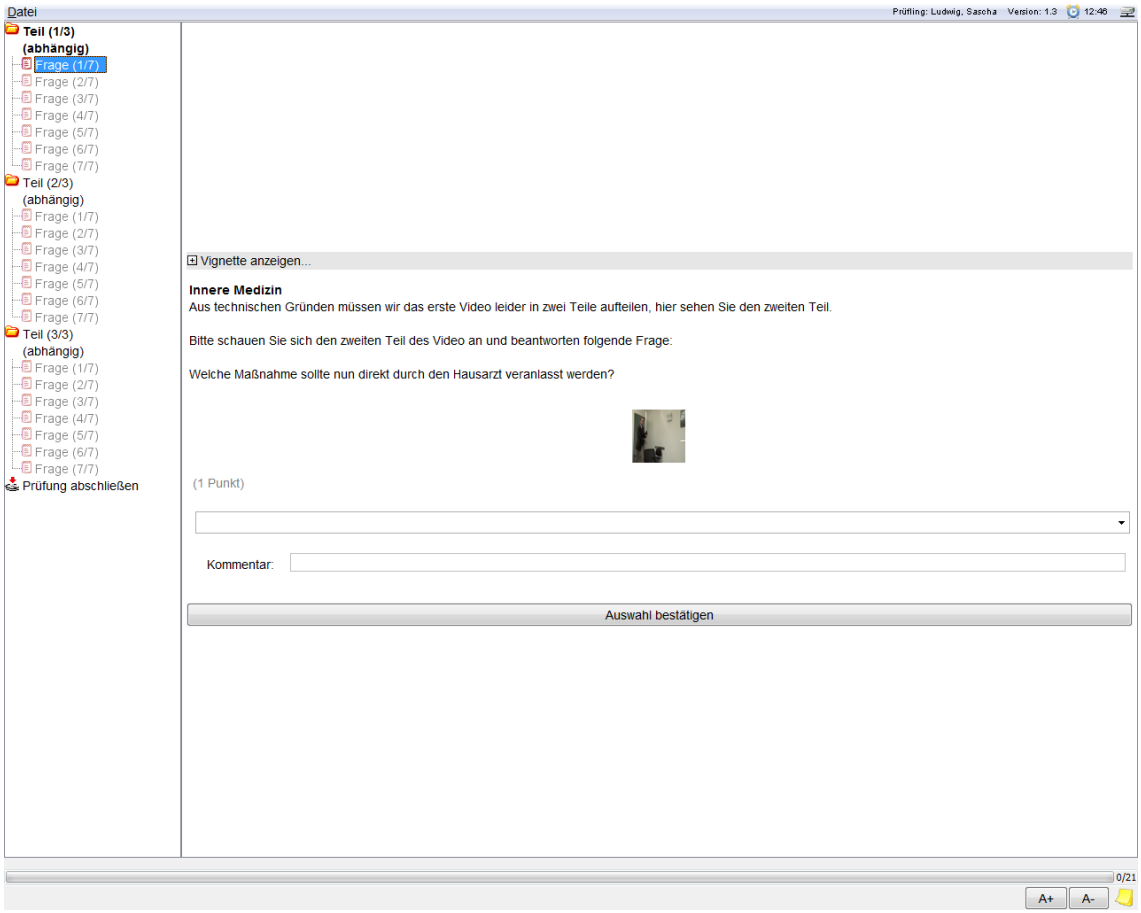


Abbildung A9: Bildschirmfoto eines Video-*Key-Feature*-Falls. *CAMPUS-Prüfungsplayer* mit der Darstellung eines Video-*Key-Feature*-Items mit Miniaturbild des Videos, welches sich nach Klick auf selbiges in einem neuen Fenster des *Windows Media Players* öffnet (vgl. Abbildung 4 auf Seite 38).

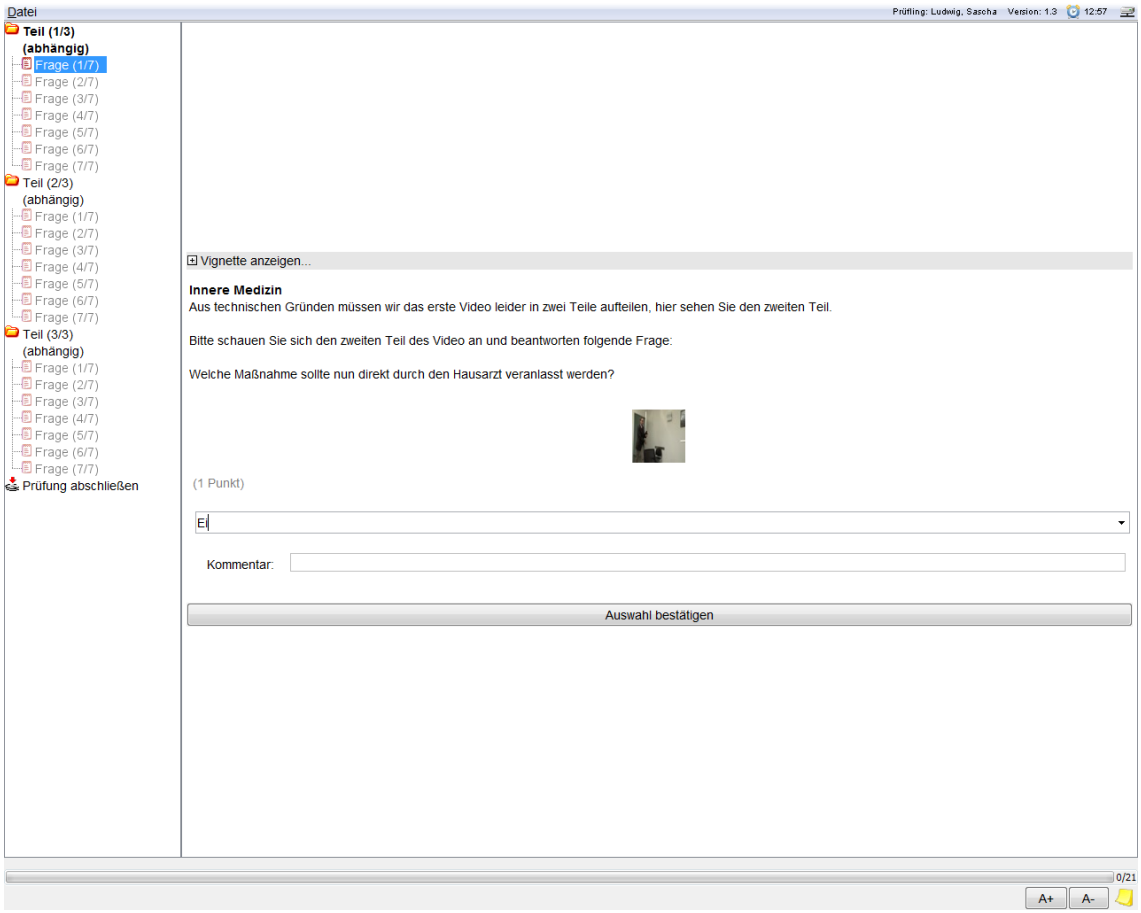


Abbildung A10: Bildschirmfoto eines Video-*Key-Feature*-Falls. *CAMPUS-Prüfungsplayer* mit der Darstellung eines Video-*Key-Feature*-Items mit Miniaturbild des Videos. Hier sind bereits zwei Buchstaben in das Eingabefeld eingetragen; mit der Eingabe des dritten Buchstabens öffnet sich ein Aufklappmenü mit den Begriffen des *Long Menus*, die diese drei Buchstaben in dieser Reihenfolge an beliebiger Stelle enthalten. Aus technischen Gründen kann hiervon kein Bildschirmfoto gemacht werden.

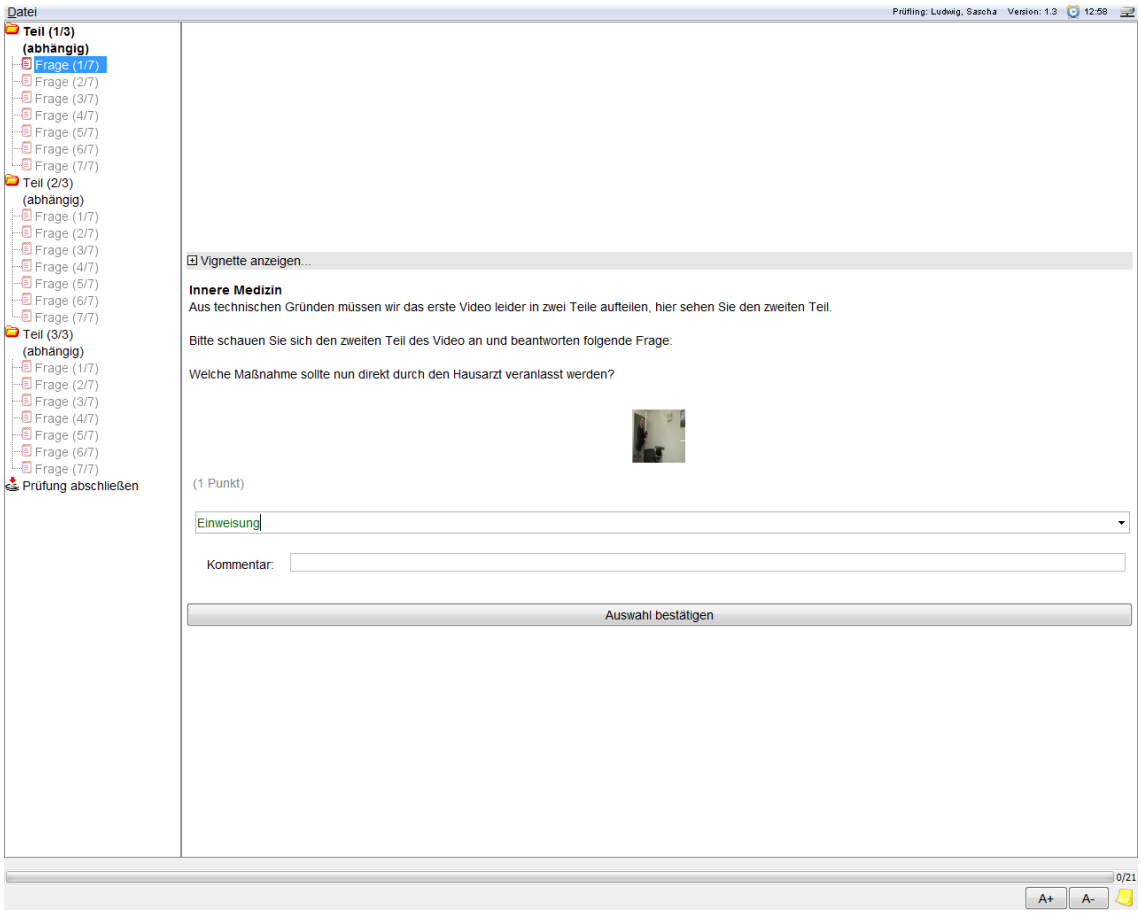


Abbildung A11: Bildschirmfoto eines Video-*Key-Feature*-Falls. *CAMPUS-Prüfungsplayer* mit der Darstellung eines Video-*Key-Feature*-Items mit Miniaturbild des Videos. Hier wurde nach Eingabe weiterer Buchstaben (insgesamt mindestens drei) bereits eine Antwort aus dem Aufklappmenü des *Long Menus* ausgewählt.

7 Literaturverzeichnis

ÄApprO 2002: Approbationsordnung für Ärzte vom 27. Juni 2002 (BGBl. I S. 2405), die zuletzt durch Artikel 3 des Gesetzes vom 16. März 2020 (BGBl. I S. 497) geändert worden ist

Altmann EM, Trafton JG, Hambrick DZ (2014): Momentary interruptions can derail the train of thought. J Exp Psychol Gen 143, 215–226

Amini M, Moghadami M, Kojuri J, Abbasi H, Abadi AAD, Molaee NA, Pishbin E, Javadzade HR, Kasmaee VM, Vakili H et al. (2011): An innovative method to assess clinical reasoning skills: Clinical reasoning tests in the second national medical science Olympiad in Iran. BMC Res Notes 4, 418

Andresen JC: Effektivität von Key-Feature-Prüfungen beim Erwerb der Kompetenz Clinical Reasoning in der medizinischen Ausbildung. Med. Diss. Göttingen 2020

Baghdady M, Carnahan H, Lam EWN, Woods NN (2014): Test-enhanced learning and its effect on comprehension and diagnostic accuracy. Med Educ 48, 181–188

Balslev T, de Grave WS, Muijtjens AMM, Scherpbier AJJA (2005): Comparison of text and video cases in a postgraduate problem-based learning format. Med Educ 39, 1086–1092

Bangert-Drowns RL, Kulik JA, Kulik CLC (1991): Effects of frequent classroom testing. J Educ Res 85, 89–99

Barrows HS (1996): Problem-based learning in medicine and beyond: A brief overview. New Dir Teach Learn 1996, 3–12

Basu Roy R, McMahon GT (2012): Video-based cases disrupt deep critical thinking in problem-based learning. Med Educ 46, 426–435

Bauer D, Holzer M, Kopp V, Fischer MR (2011): Pick-N multiple choice-exams: A comparison of scoring algorithms. Adv Heal Sci Educ 16, 211–221

Biggs JB: Teaching for Quality Learning at University: What the Student does. 3. Auflage; Open University Press, Maidenhead 2007

Bordage G, Page G: An alternate approach to PMPs, the key feature concept; in: Further developments in assessing clinical competence; hrsg. v. Hart IR, Harden R; Can-Heal Publications, Montreal 1987, 57–75

Bortz J, Döring N: Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. 4. Auflage; Springer Medizin, Heidelberg 2006

Brüstle P: Kurzanleitung Prüfen mit MC-Fragen. Studiendekanat der Medizinischen Fakultät, Albert-Ludwigs-Universität, Freiburg 2011

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

Callahan CA, Hojat M, Gonnella JS (2007): Volunteer bias in medical education research: An empirical study of over three decades of longitudinal data. Med Educ 41, 746–753

Carpenter SK (2009): Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. J Exp Psychol Learn Mem Cogn 35, 1563–1569

Carpenter SK, DeLosh EL (2006): Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. Mem Cognit 34, 268–276

Chan LK, Patil NG, Chen JY, Lam JCM, Lau CS, Ip MSM (2010): Advantages of video trigger in problem-based learning. Med Teach 32, 760–765

Cohen J (1992): A power primer. Psychol Bull 112, 155–159

Cook DA, Durning SJ, Sherbino J, Gruppen LD (2019): Management reasoning. Acad Med 94, 1310–1316

Croskerry P (2009): A universal model of diagnostic reasoning. Acad Med 84, 1022–1028

Dawson NV (1993): Physician judgment in clinical settings: Methodological influences and cognitive performance. Clin Chem 39, 1468–1480

Dienes Z: Understanding psychology as a science: An introduction to scientific and statistical inference. Palgrave Macmillan, Basingstoke 2008

Dobson JL, Linderholm T (2015): Self-testing promotes superior retention of anatomy and physiology information. Adv Health Sci Educ Theory Pract 20, 149–161

Dolmans DHJM, Gijselaers WH, Moust JHC, de Grave WS, Wolfhagen IHAP, van der Vleuten CPM (2002): Trends in research on the tutor in problem-based learning: Conclusions and implications for educational practice and research. Med Teach 24, 173–180

Dong C, Goh PS (2015): Twelve tips for the effective use of videos in medical education. Med Teach 37, 140–145

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

Farmer EA, Page G (2005): A practical guide to assessing clinical decision-making skills using the key features approach. BMC 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

George D, Mallery P: SPSS for Windows step by step: A simple guide and reference, 11.0 update. 4. Auflage; Allyn & Bacon, Boston 2003

Godden DR, Baddeley AD (1975): Context-dependent memory in two natural environments: On land and underwater. Br J Psychol 66, 325–331

Hatala R, Norman GR (2002): Adapting the key features examination for a clinical clerkship. Med Educ 36, 160–165

Herzig S, Biehl L, Stelberg H, Hick C, Schmeißer N, Koerfer A (2006): Wann ist ein Arzt ein guter Arzt? Dtsch Med Wochenschr 131, 2883–2888

Hogarth RM: Educating intuition. The University of Chicago Press, Chicago 2001

Holtzman KZ, Swanson DB, Ouyang W, Hussie K, Allbee K (2009): Use of multimedia on the step 1 and step 2 clinical knowledge components of USMLE: A controlled trial of the impact on item characteristics. Acad Med 84, 90–93

Hrynchak P, Glover Takahashi S, Nayer M (2014): Key-feature questions for assessment of clinical reasoning: A literature review. Med Educ 48, 870–883

Jacoby LL, Shimizu Y, Velanova K, Rhodes MG (2005): Age differences in depth of retrieval: Memory for foils. J Mem Lang 52, 493–504

Jeffreys H: The theory of probablility. 3. Auflage; Oxford University Press, Oxford 1998

Kamin C, O'Sullivan P, Deterding R, Younger M (2003): A comparison of critical thinking in groups of third-year medical students in text, video and virtual PBL case modalities. Acad Med 78, 204–211

Karpicke JD, Roediger HL (2007): Repeated retrieval during learning is the key to long-term retention. J Mem Lang 57, 151–162

Kassirer JP (2010): Teaching clinical reasoning: Case-based and coached. Acad Med 85, 1118–1124

Kirschbaum C, Hellhammer DH (1994): Salivary cortisol in psychoneuroendocrine research: Recent developments and applications. Psychoneuroendocrinology 19, 313–333

Koole S, Dornan T, Aper L, Scherpbier A, Valcke M, Cohen-Schotanus J, Derese A (2012): Does reflection have an effect upon case-solving abilities of undergraduate medical students? BMC Med Educ 12, 75

Kopp V, Möltner A, Fischer MR (2006): Key-Feature-Probleme zum Prüfen von prozeduralem Wissen: Ein Praxisleitfaden. GMS Z Med Ausbild 23, Doc50

Krathwohl DR (2002): A revision of Bloom's taxonomy: An overview. Theory Pract 41, 212–218

Krebs R: Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung. Institut für Medizinische Lehre, Universität Bern, Bern 2004

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

Kromann CB, Bohnstedt C, Jensen ML, Ringsted C (2010): The testing effect on skills learning might last 6 months. Adv Health Sci Educ Theory Pract 15, 395–401

Kromann CB, Jensen ML, Ringsted C (2011): Test-enhanced learning may be a gender-related phenomenon explained by changes in cortisol level. Med Educ 45, 192–199

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 (2013a): Comparative effects of test-enhanced learning and self-explanation on long-term retention. Med Educ 47, 674–682

Larsen DP, Butler AC, Lawson AL, Roediger HL (2013b): The importance of seeing the patient: Test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. Adv Health Sci Educ Theory Pract 18, 409–425

Logan JM, Thompson AJ, Marshak DW (2011): Testing to enhance retention in human anatomy. Anat Sci Educ 4, 243–248

Mayer RE, Sims VK (1994): For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. J Educ Psychol 86, 389–401

Mayer RE, Moreno R (2003): Nine ways to reduce cognitive load in multimedia learning. Educ Psychol 38, 43–52

McGuire CH, Solomon LM, Bashook PG: Construction and use of written simulations. Psychological Corporation of Harcourt, Brace, Jovanovich, New York 1976

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

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

Mousavi SY, Low R, Sweller J (1995): Reducing cognitive load by mixing auditory and visual presentation modes. J Educ Psychol 87, 319–334

NKLM 2015: Medizinischer Fakultätentag der Bundesrepublik Deutschland e. V.: Nationaler Kompetenzbasierter Lernzielkatalog Medizin (NKLM). Berlin 2015

Norman GR, Schmidt HG (1992): The psychological basis of problem-based learning: A review of the evidence. Acad Med 67, 557–565

Offener Brief „Masterplan Medizinstudium 2020“: Medizinstudierende im Hartmannbund und der bvmd: Offener Brief zum „Masterplan Medizinstudium 2020“. Berlin 2016

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

Paivio A: Imagery and verbal processes. Holt, Rinehart and Winston, New York 1971

Pluck G, Johnson H (2011): Stimulating curiosity to enhance learning. Educ Sci Psychol 2, 24–31

Pyc MA, Rawson KA (2009): Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? J Mem Lang 60, 437–447

Raupach T, Münscher C, Anders S, Steinbach R, Pukrop T, Hege I, Tullius M (2009): Web-based collaborative training of clinical reasoning: A randomized trial. Med Teach 31, e431–437

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

Raupach T, 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

Reusser K (2005): Problemorientiertes Lernen – Tiefenstruktur, Gestaltungsformen, Wirkung. Beiträge zur Lehrerinnen- und Lehrerbildung 23, 159–182

La Rochelle JS, Durning SJ, Pangaro LN, Artino AR, van der Vleuten CPM, Schuwirth LWT (2011): Authenticity of instruction and student performance: A prospective randomised trial. Med Educ 45, 807–817

La Rochelle JS, Durning SJ, Pangaro LN, Artino AR, van der Vleuten CPM, Schuwirth LWT (2012): Impact of increased authenticity in instructional format on preclerkship students’ performance: A two-year, prospective, randomized study. Acad Med 87, 1341–1347

Rodriguez MC (2005): Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educ Meas Issues Pract 24, 3–13

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

Rogausch A, Hofer R, Krebs R (2010): Rarely selected distractors in high stakes medical multiple-choice examinations and their recognition by item authors: A simulation and survey. BMC Med Educ 10, 85

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

Schuelper N, Ludwig S, Anders S, Raupach T (2019): The impact of medical students' individual teaching format choice on the learning outcome related to clinical reasoning. JMIR Med Educ 5, e13386

Schulze J, Drolshagen S (2006): Format und Durchführung schriftlicher Prüfungen. GMS Z Med Ausbild 23, Doc44

Schuwirth LWT, van der Vleuten CPM (2004): Different written assessment methods: What can be said about their strengths and weaknesses? Med Educ 38, 974–979

Schuwirth LWT, van der Vleuten CPM, Stoffers HEJH, Peperkamp AGW (1996): Computerized long-menu questions as an alternative to open-ended questions in computerized assessment. Med Educ 30, 50–55

Stanovich KE: The robot's rebellion: Dinding meaning in the age of Darwin. University of Chicago Press, Chicago 2004

Sweller J, van Merrienboer JJG, Paas FGWC (1998): Cognitive architecture and instructional design. Educ Psychol Rev 10, 251–296

Tulving E, Thomson DM (1973): Encoding specificity and retrieval processes in episodic memory. Psychol Rev 80, 352–373

van der Vleuten CPM, Newble DI (1995): How can we test clinical reasoning? Lancet 345, 1032–1034

Vogel T (2018): Ärztliche Prüfung: Das Stex im Wandel. Dtsch Arztebl Med Stud WS2018/19, 7–9

Vohs KD, Baumeister RF, Loewenstein G (Hrsg.): Do emotions help or hurt decision making? A hedgefoxian perspective. Russell Sage Foundation, New York 2007

Weih M, Harms D, Rauch C, Segarra L, Reulbach U, Degirmenci U, de Zwaan M, Schwab S, Kornhuber J (2009): Qualitätsverbesserung von Multiple-Choice-Prüfungen. Nervenarzt 80, 324–328

Wetzels R, Matzke D, Lee MD, Rouder JN, Iverson GJ, Wagenmakers EJ (2011): Statistical evidence in experimental psychology. Perspect Psychol Sci 6, 291–298

Wing EA, Marsh EJ, Cabeza R (2013): Neural correlates of retrieval-based memory enhancement: An fMRI study of the testing effect. Neuropsychologia 51, 2360–2370

Woodham LA, Ellaway RH, Round J, Vaughan S, Poulton T, Zary N (2015): Medical student and tutor perceptions of video versus text in an interactive online virtual patient for problem-based learning: A pilot study. J Med Internet Res 17, e151

Danksagung

An dieser Stelle möchte ich mich ganz herzlich bei meinem Doktorvater und Betreuer Prof. Dr. Tobias Raupach für die Überlassung des Dissertationsthemas und die intensive und engagierte Betreuung und Unterstützung, die ich stets erhalten habe, bedanken. Nicht zuletzt durch die Präsentation auf wissenschaftlichen Konferenzen konnte ich im Rahmen dieser Arbeit wertvolle Erfahrungen sammeln. Ein besonderer Dank gilt Dr. Nikolai Schuelper für seine Unterstützung, zahlreichen Anregungen und Ratschläge, die wesentlich zum Gelingen dieser Arbeit beitrugen. Außerdem danke ich Herrn Prof. Dr. Jamie Brown für die Hilfe bei der Analyse nach Bayes und Prof. Dr. Anders für die Anregungen zum Studiendesign.

Ein besonders großer Dank geht an alle fast einhundert Schauspieler aus dem Familien- und Freundeskreis, der Theatergruppe „Lehrgut“ und dem „Theater im OP“, den Schauspielpatienten sowie den ärztlichen und nichtärztlichen Kollegen der UMG und darüber hinaus allen, die bereit waren, sich im Rahmen dieses Projektes vor die Kamera zu stellen und hinter der Kamera zu helfen. Ebenfalls danken möchte ich allen, die neben der UMG für dieses Projekt Videoausstattung (Videoteam der Niedersächsischen Staats- und Universitätsbibliothek Göttingen), Drehorte (Praxen Dr. Hübner in Detmold, Dr. Lenger in Lemgo, Dr. Köbrich in Göttingen, Nieren-Rheuma-Zentrum Standort Göttingen) und Requisiten (DRK Ortsvereine Göttingen und Lage-Lippe, Nüsse Orthopädie-Technik Göttingen) zur Verfügung gestellt haben. Ohne all diese Menschen wäre das Projekt in dieser Form nicht möglich gewesen!

Stellvertretend für das gesamte Team der Lehr-IT des Bereichs Medizindidaktik und Ausbildungsforschung der UMG möchte ich Christian Münscher für die technische Unterstützung bei der Implementierung, Vorbereitung und Durchführung der E-Fallseminare danken. Nicht zuletzt gilt mein Dank allen aktuellen und ehemaligen Mitarbeitern sowie Doktoranden des Bereichs Medizindidaktik und Ausbildungsforschung der UMG für ihre Unterstützung und Hilfsbereitschaft bei der Erstellung der Videos, Durchführung der E-Fallseminare sowie den konstruktiven Austausch, nicht nur während der Teamtreffen. Abschließend danke ich allen Medizinstudierenden der UMG für ihre Teilnahme an dieser Studie.